

# A Geometrical Study of Matching Pursuit Parametrization

October 10, 2018

## Abstract

This paper studies the effect of discretizing the parametrization of a dictionary used for Matching Pursuit decompositions of signals. Our approach relies on viewing the continuously parametrized dictionary as an embedded manifold in the signal space on which the tools of differential (Riemannian) geometry can be applied. The main contribution of this paper is twofold. First, we prove that if a discrete dictionary reaches a minimal density criterion, then the corresponding discrete MP (dMP) is equivalent in terms of convergence to a weakened hypothetical continuous MP. Interestingly, the corresponding weakness factor depends on a density measure of the discrete dictionary. Second, we show that the insertion of a simple geometric gradient ascent optimization on the atom dMP selection maintains the previous comparison but with a weakness factor at least two times closer to unity than without optimization. Finally, we present numerical experiments confirming our theoretical predictions for decomposition of signals and images on regular discretizations of dictionary parametrizations.

*Keywords:* Matching Pursuit, Riemannian geometry, Optimization, Convergence, Dictionary, Parametrization.

## 1 Introduction

There has been a large effort in the last decade to develop analysis techniques that decompose non-stationary signals into elementary components, called *atoms*, that characterize their salient features [1–5]. In particular, the matching pursuit (MP) algorithm has been extensively studied [2, 6–11] to expand a signal over a redundant dictionary of elementary atoms, based on a greedy process that selects the elementary function that best matches the residual signal at each iteration. Hence, MP progressively isolates the structures of the signal that are coherent with respect to the chosen dictionary, and provides an adaptive signal representation in which the more significant coefficients are first extracted. The progressive nature of MP is a key issue for adaptive and scalable communication applications [12, 13].

A majority of works that have considered MP for practical signal approximation and compression define the dictionary based on the discretization of a parametrized prototype function, typically a scaled/modulated Gaussian function or its second derivative [6, 14, 15]. An orthogonal 1-D or 2-D wavelet basis is also a trivial example of such a discretization even if in that case MP is not required to find signal coefficients; a simple wavelet decomposition is computationally more efficient. Works that do not directly rely on a prototype function either approximate

such a parametrized dictionary based on computationally efficient cascades of filters [16–18], or attempt to adapt a set of parametrized dictionary elements to a set of training signal samples based on vector quantization techniques [19, 20]. Thus, most earlier works define their dictionary by discretizing, directly or indirectly, the parameters of a prototype function.

The key question is then: *how should the continuous parameter space be discretized?* A fine discretization results in a large dictionary which approximates signals efficiently with few atoms, but costs both in terms of computational complexity and atom index entropy coding. Previous works have studied this trade-off empirically [6, 15]. In contrast, our paper focuses on this question in a formal way. It provides a first attempt to quantify analytically how the MP convergence is affected by the discretization of the continuous space of dictionary function parameters.

Our compass to reach this objective is the natural geometry of the continuous dictionary. This dictionary can be seen as a parametric (Riemannian) manifold on which the tools of differential geometry can be applied. This geometrical approach, of increasing interest in the signal processing literature, is inspired by the works [21, 22] on *Image Appearance Manifolds*, and is also closely linked to manifolds of parametric probability density function associated to the Fisher information metric [23]. Some preliminary hints were also provided in a Riemannian study of generalized correlation of signals with probing functions [24].

The outcome of our study is twofold. On the one hand, we analyze how the rate of convergence of the continuous MP (cMP) is affected by the discretization of the prototype function parameters. We demonstrate that the MP using that discretized dictionary (dMP) converges like a weak continuous MP, i.e. a MP algorithm where the coefficient of the selected atom at each iteration overtakes only a percentage (the *weakness factor*) of the largest atom magnitude. We describe then how this weakness factor decreases as the so-called *density radius*<sup>1</sup> of the discretization increases. This observation is demonstrated experimentally on images and randomly generated 1-D signals.

On the other hand, to improve the rate of convergence of discrete MP without resorting to a finer but computationally heavier discretization, we propose to exploit a geometric gradient ascent method. This allows to converge to a set of locally optimal continuous parameters, starting from the best set of parameters identified by a coarse but computationally light discrete MP. Each atom of the MP expansion is then defined in two steps. The first step selects the discrete set of parameters that maximizes the inner product between the corresponding dictionary function and the residual signal. The second step implements a (manifold<sup>2</sup>) gradient ascent method to compute the prototype function parameters that maximize the inner product function over the continuous parameter space. As a main analytical result, we demonstrate that this geometrically optimized discrete MP (gMP) is again equivalent to a continuous MP, but with a weakness factor that is two times closer to unity than for the non-optimized dMP. Our experiments confirm that the proposed gradient ascent procedure significantly increases the rate of convergence of MP, compared to the non-optimized discrete MP. At an equivalent convergence rate, the optimization allows reduction of the discretization density by an order of magnitude, resulting in significant computational gains.

The paper is organized as follows. In Section 2, we introduce the notions of parametric

---

<sup>1</sup>This density radius represents the maximal distance between any atom of the continuous dictionary and its closest atom in the discretization.

<sup>2</sup>In the sense that this gradient ascent evolves on the manifold induced by the intrinsic dictionary geometry.

dictionary in the context of signal decomposition in an abstract Hilbert space. This dictionary is then envisioned as a Hilbert manifold, and we describe how its geometrical structure influences its parametrization using the tools of differential geometry. Section 3 surveys the definition of (weak) continuous MP providing a theoretical optimal rate of convergence for further comparisons with other greedy decompositions. A “discretization autopsy” of this algorithm is performed in Section 4 and a resulting theorem explaining the dependences of the dMP convergence relatively to this sampling is proved. A simple but illustrative example of a 1-D dictionary, the wavelet (affine) dictionary, is then given. The optimization scheme announced above is developed in Section 5. After a review of gradient ascent optimization evolving on manifolds, the geometrically optimized MP is introduced and its theoretical rate of convergence analyzed in a second theorem. Finally, in Section 6, experiments are performed for 1-D and 2-D signal decompositions using dMP and gMP on various regular discretizations of dictionary parametrizations. We provide links to previous related works in Section 7 and conclude with possible extensions in Section 8.

## 2 Dictionary, Parametrization and Differential Geometry

Our object of interest throughout this paper is a general real “signal”, i.e. a real function  $f$  taking value on a measure space  $X$ . More precisely, we assume  $f$  in the set of finite energy signals, i.e.  $f \in L^2(X, d\mu) = \{u : X \rightarrow \mathbb{R} : \|u\|^2 = \int_X |u(x)|^2 d\mu(x) < \infty\}$ , for a certain integral measure  $d\mu(x)$ . Of course, the natural comparison of two functions  $u$  and  $v$  in  $L^2(X, d\mu)$  is realized through the scalar product  $\langle u, v \rangle_{L^2(X)} = \langle u, v \rangle \triangleq \int_X u(x)v(x)d\mu(x)$  making  $L^2(X, d\mu)$  a Hilbert<sup>3</sup> space where  $\|u\|^2 = \langle u, u \rangle$ .

This very general framework can be specialized to 1-D signal or image decomposition where  $X$  is given respectively by  $\mathbb{R}$  or  $\mathbb{R}^2$ , but also to more special spaces like the two dimensional sphere  $S^2$  [25] or the hyperboloid [26]. In the sequel, we will write simply  $L^2(X) = L^2(X, d\mu)$ .

In the following sections, we will *decompose*  $f$  over a highly redundant parametric *dictionary* of real *atoms*. These are obtained from smooth transformations of a real mother function  $g \in L^2(X)$  of unit norm. Formally, each atom is a function  $g_\lambda(x) = [U(\lambda)g](x) \in L^2(X)$ , for a certain isometric operator  $U$  parametrized by elements  $\lambda \in \Lambda$  and such that  $\|g_\lambda\| = \|g\| = 1$ . The *parametrization* set  $\Lambda$  is a continuous space where each  $\lambda \in \Lambda$  corresponds to  $P$  continuous components  $\lambda = \{\lambda^i\}_{0 \leq i \leq P-1}$  of different nature. For instance, in the case of 1-D signal or image analysis,  $g$  may be transformed by translation, modulation, rotation, or (anisotropic) dilation operations, each associated to one component  $\lambda^i$  of  $\lambda$ . Our dictionary is then the set  $\text{dict}(g, U, \Theta) \triangleq \{g_\lambda(x) = [U(\lambda)g](x) : \lambda \in \Theta\}$ , for a certain subset  $\Theta \subseteq \Lambda$ . In the rest of the paper, we write  $\text{dict}(\Theta) = \text{dict}(g, U, \Theta)$ , assuming  $g$  and  $U$  implicitly given by the context. For the case  $\Theta = \Lambda$ , we write  $\mathcal{D} = \text{dict}(\Lambda)$ .

We assume that  $g$  is twice differentiable over  $X$  and that the functions  $g_\lambda(x)$  are twice differentiable on each of the  $P$  components of  $\lambda$ . In the following, we write  $\partial_i$  for the partial derivative with respect to  $\lambda^i$ , i.e.  $\frac{\partial}{\partial \lambda^i}$ , of any element (e.g.  $g_\lambda(x)$ ,  $\langle g_\lambda, u \rangle$ , ...) depending on  $\lambda$ , and  $\partial_{ij} = \partial_i \partial_j$ . From the smoothness of  $U$  and  $g$ , we have  $\partial_{ij} = \partial_{ji}$  on quantities built from these two ingredients.

Let us now analyze the geometrical structure of  $\Lambda$ . Rather than an artificial *Euclidean distance*

---

<sup>3</sup>Assuming it *complete*, i.e. every Cauchy sequence converges in this space relatively to the norm  $\|\cdot\|^2 = \langle \cdot, \cdot \rangle$ .

$d_{\mathcal{E}}(\lambda_a, \lambda_b)^2 \triangleq \sum_i (\lambda_a^i - \lambda_b^i)^2$  between  $\lambda_a, \lambda_b \in \Lambda$ , we use a distance introduced by the dictionary  $\mathcal{D}$  itself seen as a  $P$ -dimensional parametric submanifold of  $L^2(X)$  (or a *Hilbert manifold*<sup>4</sup> [27]). The *dictionary distance*  $d_{\mathcal{D}}$  is thus the distance in the embedding space  $L^2(X)$ , i.e.  $d_{\mathcal{D}}(\lambda_a, \lambda_b) \triangleq \|g_{\lambda_a} - g_{\lambda_b}\|$ .

From this embedding, we can define an intrinsic distance in  $\mathcal{D}$ , namely the *geodesic distance*. This later has been used in a similar context in the work of Grimes and Donoho [22] and we follow here their approach. For our two points  $\lambda_a, \lambda_b$ , assume that we have a smooth curve  $\gamma : [0, 1] \rightarrow \Lambda$  with  $\gamma(t) = (\gamma^0(t), \dots, \gamma^{P-1}(t))$ , such that  $\gamma(0) = \lambda_a$  and  $\gamma(1) = \lambda_b$ . The length  $\mathcal{L}(\gamma)$  of this curve in  $\mathcal{D}$  is thus given by  $\mathcal{L}(\gamma) \triangleq \int_0^1 \|\frac{d}{dt} g_{\gamma(t)}\| dt$ , assuming that  $g_{\gamma(t)}$  is differentiable<sup>5</sup> with respect to  $t$ .

The *geodesic distance* between  $\lambda_a$  and  $\lambda_b$  in  $\Lambda$  is the length of shortest path between these two points, i.e.

$$d_{\mathcal{G}}(\lambda_a, \lambda_b) \triangleq \inf_{\gamma(\lambda_a \rightarrow \lambda_b)} \int_0^1 \|\frac{d}{dt} g_{\gamma(t)}\| dt, \quad (1)$$

where  $\gamma(\lambda_a \rightarrow \lambda_b)$  is any differentiable curve  $\gamma(t)$  linking  $\lambda_a$  to  $\lambda_b$  for  $t$  equals to 0 and 1 respectively.

We denote by  $\gamma_{\lambda_a \lambda_b}$  the optimal *geodesic* curve joining  $\lambda_a$  and  $\lambda_b$  on the manifold  $\mathcal{D}$ , i.e. such that  $\mathcal{L}(\gamma_{\lambda_a \lambda_b}) = d_{\mathcal{G}}(\lambda_a, \lambda_b)$ , and we assume henceforth that it is always possible to define this curve between two points of  $\Lambda$ . Note that by construction,  $d_{\mathcal{G}}(\lambda_a, \lambda_b) = d_{\mathcal{G}}(\lambda_a, \lambda') + d_{\mathcal{G}}(\lambda', \lambda_b)$ , for all  $\lambda'$  on the curve  $\gamma_{\lambda_a \lambda_b}(t)$ .

In the language of differential geometry, the parameter space  $\Lambda$  is a Riemannian manifold  $\mathcal{M} = (\Lambda, \mathcal{G}_{ij})$  with metric  $\mathcal{G}_{ij}(\lambda) = \langle \partial_i g_{\lambda}, \partial_j g_{\lambda} \rangle$ . Indeed, for any differentiable curve  $\gamma : t \in [-\delta, \delta] \rightarrow \gamma(t) \in \Lambda$  with  $\delta > 0$  and  $\gamma(0) = \lambda$ , we have

$$\|\frac{d}{dt} g_{\gamma(t)}|_{t=0}\|^2 = \dot{\gamma}^i(0) \dot{\gamma}^j(0) \mathcal{G}_{ij}(\lambda), \quad (2)$$

with  $\dot{u}(t) = \frac{d}{dt} u(t)$ , and where Einstein's summation convention is used for simplicity<sup>6</sup>.

The vector  $\xi^i = \dot{\gamma}^i(0)$  is by definition a vector in the tangent space  $T_{\lambda}\Lambda$  of  $\Lambda$  in  $\lambda$ . The meaning of relation (2) is that the metric  $\mathcal{G}_{ij}(\lambda)$  allows the definitions of a scalar product and a norm in each  $T_{\lambda}\Lambda$ . The norm of a vector  $\xi \in T_{\lambda}\Lambda$  is therefore noted  $|\xi|^2 = |\xi|_{\lambda}^2 \triangleq \xi^i \xi^j \mathcal{G}_{ij}(\lambda)$ , with the correspondence  $\|\frac{d}{dt} g_{\gamma(t)}|_{t=0}\| = |\dot{\gamma}|$ . For the consistency of further Riemannian geometry developments, we assume that our dictionary  $\mathcal{D}$  is *non-degenerate*, i.e. that it induces a positive definite metric  $\mathcal{G}_{ij}$ . Appendix A provides additional details.

We conclude this section with the *arc length* (or *curvilinear*) parametrization “ $s$ ” [28] of a curve  $\gamma(s)$ . It is such that  $|\gamma'|^2 \triangleq \gamma'^i(s) \gamma'^j(s) \mathcal{G}_{ij}(\gamma(s)) = 1$ , where  $u'(s) = \frac{d}{ds} u(s)$ . From its definition, the curvilinear parameter  $s$  is the one which measures at each point  $\gamma(s)$  the length of the segment of curve already travelled on  $\gamma$  from  $\gamma(0)$ . Therefore, in this parametrization,  $\lambda_a = \gamma_{\lambda_a \lambda_b}(0)$  and  $\lambda_b = \gamma_{\lambda_a \lambda_b}(d_{\mathcal{G}}(\lambda_a, \lambda_b))$ .

<sup>4</sup>This is a special case of Image Appearance Manifold (IAM) defined for instance in [21, 22]. It is also closely linked to manifolds of parametric probability density function associated to the Fisher information metric [23].

<sup>5</sup>Another definition of  $\mathcal{L}$  exists for non differentiable curve. See for instance [22].

<sup>6</sup>Namely, a summation in an expression is defined implicitly each time the same index is repeated once as a subscript and once as a superscript, the range of summation being always  $[0, P-1]$ , so that for instance the expression  $a^i b_i$  reads  $\sum_{i=0}^{P-1} a^i b_i$ .

### 3 Matching Pursuit in Continuous Dictionary

Let us assume that we want to decompose a function  $f \in L^2(X)$  into simpler elements (atoms) coming from a dictionary  $\text{dict}(\Theta)$ , given a possibly uncountable and infinite subset  $\Theta \subseteq \Lambda$ . Our general aim is thus to find a set of *coefficients*  $\{c_m\}$  such that  $f(x)$  is equal or well approximated by  $f_{\text{app}}(x) = \sum_m c_m g_{\lambda_m}(x)$  with a finite set of atoms  $\{g_{\lambda_m}\} \subset \text{dict}(\Theta)$ .

Formally, for a given *weakness* factor  $\alpha \in (0, 1]$ , a *General Weak( $\alpha$ ) Matching Pursuit* decomposition of  $f$  [2, 29], written  $\text{MP}(\Theta, \alpha)$ , in the dictionary  $\text{dict}(\Theta)$  is performed through the following *greedy*<sup>7</sup> algorithm :

$$R^0 f = f, \quad A^0 f = 0, \quad (\text{initialization}),$$

$$R^{m+1} f = R^m f - \langle g_{\lambda_{m+1}}, R^m f \rangle g_{\lambda_{m+1}}, \quad (3a)$$

$$A^{m+1} f = A^m f + \langle g_{\lambda_{m+1}}, R^m f \rangle g_{\lambda_{m+1}}, \quad (3b)$$

$$\text{with : } \langle g_{\lambda_{m+1}}, R^m f \rangle^2 \geq \alpha^2 \sup_{\lambda \in \Theta} \langle g_{\lambda}, R^m f \rangle^2. \quad (3c)$$

The quantity  $R^{m+1} f$  is the *residual* of  $f$  at iteration  $m+1$ . Since it is orthogonal to atom  $g_{\lambda_{m+1}}$ ,  $\|R^{m+1} f\|^2 = \|R^m f\|^2 - \langle g_{\lambda_{m+1}}, R^m f \rangle^2 \leq \|R^m f\|^2$ , so that the energy  $\|R^m f\|^2$  is non-increasing. The function  $A^m f$  is the *m-term approximation* of  $f$  with  $A^m f = \sum_{k=0}^{m-1} \langle g_{\lambda_{k+1}}, R^k f \rangle g_{\lambda_{k+1}}$ .

Notice that the *selection rule* (3c) concerns the square of the real scalar product  $\langle g_{\lambda}, R^m f \rangle$ . Matching Pursuit atom selection is typically defined over the absolute value  $|\langle g_{\lambda}, R^m f \rangle|$ . However, we prefer this equivalent quadratic formulation first to avoid the abrupt behavior of the absolute value when the scalar product crosses zero, and second for consistency with the quadratic optimization framework to be explained in Section 5. Finally, to allow the non-weak case where  $\alpha = 1$ , we assume that a maximizer  $g_u \in \text{dict}(\Theta)$  of  $\langle g, u \rangle^2$  always exists for any  $u \in L^2(X)$ .

If  $\Theta$  is uncountable, our general Matching Pursuit algorithm is named *continuous Matching pursuit*. In particular, for  $\Theta = \Lambda$ , we write  $\text{cMP}(\alpha) = \text{MP}(\Lambda, \alpha)$ . The *rate of convergence* (or convergence) of the  $\text{cMP}(\alpha)$ , characterized by the rate of decay of  $\|R^m f\|$  with  $m$ , can be assessed in certain particular cases. For instance, if there exists a Hilbert space  $\mathcal{S} \subseteq L^2(X)$  containing  $\mathcal{D} = \text{dict}(\Lambda)$  such that

$$\beta^2 = \inf_{u \in \mathcal{S}, \|u\|=1} \sup_{\lambda \in \Lambda} \langle g_{\lambda}, u \rangle^2 > 0, \quad (4)$$

then the  $\text{cMP}(\alpha)$  converges inside  $\mathcal{S}$ . In fact, the convergence is exponential [30] since  $\langle g_{\lambda_m}, R^{m-1} f \rangle^2 \geq \alpha^2 \beta^2 \|R^{m-1} f\|^2$  and  $\|R^m f\|^2 \leq \|R^{m-1} f\|^2 - \alpha^2 \beta^2 \|R^{m-1} f\|^2 \leq (1 - \alpha^2 \beta^2)^m \|f\|^2$ . We name  $\beta = \beta(\mathcal{S}, \mathcal{D})$  the *greedy factor* since it characterizes the MP convergence (greediness).

The existence of the greedy factor  $\beta$  is obvious for instance for finite dimensional space [30], i.e.  $f \in \mathbb{C}^N$ , with finite dictionary (finite number of atoms).

For a finite dictionary in an infinite dimensional space, as  $L^2(X)$ , the existence of  $\beta$  is not guaranteed over the whole space. However, there exists on the space of functions given by linear combination of dictionary elements, the number of terms being restricted by the dictionary (*cumulative*) *coherence* [29].

In the case of an infinite dictionary in an infinite dimension space where the greedy factor vanishes,  $\text{cMP}(\alpha)$  convergence is characterized differently on the subspace of linear combination

---

<sup>7</sup>Greedy in the sense that it does not solve a global  $\ell_0$  or  $\ell_1$  minimization [1] to find the coefficients  $c_m$  of  $f_{\text{app}}$  above, but works iteratively by solving at each iteration step a local and smaller minimization problem.

of countable subsets of dictionary elements. This question is addressed separately in a companion Technical Report [31] to this article. We now consider only the case where a non-zero greedy factor exists to characterize the rate of convergence of MP using continuous and discrete dictionaries.

## 4 Discretization effects of Continuous Dictionary

The greedy algorithm  $\text{cMP}(\alpha)$  using the dictionary  $\mathcal{D}$  is obviously numerically unachievable because of the intrinsic continuity of its main ingredient, namely the parameter space  $\Lambda$ . Any computer implementation needs at least to discretize the parametrization of the dictionary, more or less densely, leading to a countable set  $\Lambda_d \subset \Lambda$ . This new parameter space leads naturally to the definition of a countable subdictionary  $\mathcal{D}_d = \text{dict}(\Lambda_d)$ . Henceforth, elements of  $\Lambda_d$  are labelled with roman letters, e.g.  $k$ , to distinguish them from the continuous greek-labeled elements of  $\Lambda$ , e.g.  $\lambda$ .

For a weakness factor  $\alpha \in (0, 1]$ , the *discrete Weak( $\alpha_d$ ) Matching Pursuit* algorithm, or  $\text{dMP}(\alpha)$ , of a function  $f \in L^2(X)$  over  $\mathcal{D}_d$  is naturally defined as  $\text{dMP}(\alpha) = \text{MP}(\Lambda_d, \alpha)$ . The replacement of  $\Lambda$  by  $\Lambda_d$  in the MP algorithm (3) leads obviously to the following question that we address in the next section.

**Question 1.** *How does the MP rate of convergence evolve when the parametrization of a dictionary is discretized and what are the quantities that control (or bound) this evolution ?*

### 4.1 Discretization Autopsy

By working with  $\mathcal{D}_d$  instead of  $\mathcal{D}$ , the atoms selected at each iteration of  $\text{dMP}(\alpha)$  are of course less optimal than those available in the continuous framework. Answering Question 1 requires a quantitative measure of the induced loss in the MP coefficients. More concretely, defining the *score function*  $S_u(\lambda) = \langle g_\lambda, u \rangle^2$  for some  $u \in L^2(X)$ , we must analyze the difference between a maximum of  $S_u$  computed over  $\Lambda$  and that obtained from  $\Lambda_d$ . This function  $u$  will be next identified with the residue of  $\text{dMP}(\alpha)$  at any iteration to characterize the global change in convergence.

We propose to found our analysis on the geometric tools described in Section 2.

**Definition 1.** *The value  $S_u(\lambda_a)$  is critical in the direction of  $\lambda_b$  if, given the geodesic  $\gamma = \gamma_{\lambda_a \lambda_b}$  in the manifold  $\mathcal{M} = (\Lambda, \mathcal{G}_{ij})$ ,  $\frac{d}{ds} S_u(\gamma(s))|_{s=0} = 0$ , where  $\gamma(0) = \lambda_a$ .*

Notice that if  $S_u(\lambda_a)$  is critical in the direction of  $\lambda_b$ ,  $\gamma'^i(0) \partial_i S_u(\lambda_a) = 0$ . An *umbilical* point for which  $\partial_i S_u(\lambda_a) = 0$  for all  $i$ , is obviously critical in any direction. An umbilical point corresponds geometrically either to maxima, minima or saddlepoints of  $S_u$  relatively to  $\Lambda$ .

**Proposition 1.** *Given  $u \in L^2(X)$ , if  $S_u(\lambda_a)$  is critical in the direction of  $\lambda_b$  for  $\lambda_a, \lambda_b \in \Lambda$ , then for some  $r \in (0, d_G(\lambda_a, \lambda_b))$ ,*

$$|S_u(\lambda_a) - S_u(\lambda_b)| \leq \|u\|^2 d_G(\lambda_a, \lambda_b)^2 \left(1 + \left\| \frac{d^2 g_\gamma}{ds^2} \right\|_{s=r} \right), \quad (5)$$

where  $\gamma(s) = \gamma_{\lambda_a \lambda_b}(s)$  is the geodesic in  $\mathcal{M}$  linking  $\lambda_a$  to  $\lambda_b$ .



*Proof.* Let us define the twice differentiable function  $\psi(s) \triangleq S_u(\gamma(s))$  on  $s \in [0, \eta]$ , with  $\eta \triangleq d_{\mathcal{G}}(\lambda_a, \lambda_b)$ . A second order Taylor development of  $\psi$  gives, for a certain  $r \in (0, s)$ ,  $\psi(s) = \psi(0) + s\psi'(0) + \frac{1}{2}s^2\psi''(r)$ . Since  $\psi'(0) = \gamma'^i(0)\partial_i S_u(\lambda_a) = 0$  by hypothesis, we have in  $s = \eta$ ,  $|\psi(0) - \psi(\eta)| = |S_u(\lambda_a) - S_u(\lambda_b)| \leq \frac{1}{2}\eta^2 |\psi''(r)|$ . However, on any  $s$ ,  $|\psi''(s)| = 2|\langle \frac{d}{ds}g_{\gamma(s)}, u \rangle|^2 + \langle g_{\gamma(s)}, u \rangle \langle \frac{d^2}{ds^2}g_{\gamma(s)}, u \rangle| \leq 2(\|\frac{d}{ds}g_{\gamma(s)}\|^2 + \|\frac{d^2}{ds^2}g_{\gamma(s)}\|)\|u\|^2$ , using the Cauchy-Schwarz (CS) inequality in  $L^2(X)$  in the last equation. The result follows from the fact that  $\|\frac{d}{ds}g_{\gamma(s)}\| = 1$ .  $\square$

The previous Lemma is particularly important since it bounds the loss in coefficient value when we decide to choose  $S_u(\lambda_b)$  instead of the optimal  $S_u(\lambda_a)$  in function of the geodesic distance  $d_{\mathcal{G}}(\lambda_a, \lambda_b)$  between the two parameters. To obtain a more satisfactory control of this difference, we need however a new property of the dictionary.

We start by defining the *principal curvature* in the point  $\lambda \in \Lambda$  as

$$\mathcal{K}_\lambda \triangleq \sup_{\xi: \|\xi\|=1} \left\| \frac{d^2}{ds^2} g_{\gamma_\xi(s)} \Big|_{s=0} \right\|, \quad (6)$$

where  $\gamma_\xi$  is the unique geodesic in  $\mathcal{M}$  starting from  $\lambda = \gamma_\xi(0)$  and with  $\gamma'_\xi(0) = \xi$ , for a direction  $\xi$  of unit norm in  $T_\lambda\Lambda$ .

**Definition 2.** The condition number of a dictionary  $\mathcal{D}$  is the number  $\mathcal{K}^{-1}$  obtained from

$$\mathcal{K} \triangleq \sup_{\lambda \in \Lambda} \mathcal{K}_\lambda. \quad (7)$$

If  $\mathcal{K}$  does not exist (not bounded  $\mathcal{K}_\lambda$ ), by extension,  $\mathcal{D}$  is said to be of zero condition number.

The notion of condition number has been introduced by Niyogi et al. [32] to bound the local curvature of an embedded manifold<sup>8</sup> in its ambient space, and to characterize its self-avoidance. Essentially, it is the inverse of the maximum radius of a sphere that, when placed tangent to the manifold at any point, intersects the manifold only at that point [33, 34]. Our quantity  $\mathcal{K}^{-1}$  is then by construction a similar notion for the dictionary  $\mathcal{D}$  seen as a manifold in  $L^2(X)$ . However, it does not actually prevent manifold self-crossing on large distance due to the locality of our differential analysis<sup>9</sup>.

**Proposition 2.** For a dictionary  $\mathcal{D} = \text{dict}(\Lambda)$ ,

$$1 \leq \mathcal{K} \leq \sup_{\lambda \in \Lambda} \left[ \langle \partial_{ij} g_\lambda, \partial_{kl} g_\lambda \rangle \mathcal{G}^{ik} \mathcal{G}^{jl} \right]^{\frac{1}{2}}, \quad (8)$$

where  $\mathcal{G}^{ij} = \mathcal{G}^{ij}(\lambda)$  is the inverse<sup>10</sup> of  $\mathcal{G}_{ij}$ .

<sup>8</sup>In their work, the condition number, named there  $\tau^{-1}$ , of a manifold  $\mathcal{M}'$  measures the maximal “thickness”  $\tau$  of the *normal bundle*, the union of all the orthogonal complement of every tangent plane at every point of the manifold.

<sup>9</sup>A careful study of local self-avoidance of well-conditioned dictionary would have to be considered but this is beyond the scope of this paper.

<sup>10</sup>Using Einstein convention, this means  $\mathcal{G}^{ik}\mathcal{G}_{kj} = \mathcal{G}_{jk}\mathcal{G}^{ki} = \delta_j^i$ , for the Kronecker’s symbol  $\delta_j^i = \delta_{ij} = \delta^{ij} = 1$  if  $i = j$  and 0 if  $i \neq j$ .

The proof is given in Appendix B since it uses some elements of differential geometry not essential in the core of this paper. The interested reader will find also there a slightly lower bound than the bound presented in (8), exploiting covariant derivatives, Laplace-Beltrami operator and scalar curvature of  $\mathcal{M}$  [28]. We can state now the following corollary of Proposition 1.

**Corollary 1.** *In the conditions of Proposition 1, if  $\mathcal{D}$  has a non-zero condition number  $\mathcal{K}^{-1}$ , then*

$$|S_u(\lambda_a) - S_u(\lambda_b)| \leq \|u\|^2 d_{\mathcal{G}}(\lambda_a, \lambda_b)^2 (1 + \mathcal{K}). \quad (9)$$

Therefore, in the  $\text{dMP}(\alpha)$  decomposition of  $f$  based on  $\mathcal{D}_d$ , even if at each iteration the exact position of the continuous optimal atom of  $\mathcal{D}$  is not known, we are now able to estimate the convergence rate of this MP provided we introduce a new quantity characterizing the set  $\Lambda_d$ .

**Definition 3.** *The density radius  $\rho_d$  of a countable parameter space  $\Lambda_d \subset \Lambda$  is the value*

$$\rho_d = \sup_{\lambda \in \Lambda} \inf_{k \in \Lambda_d} d_{\mathcal{G}}(\lambda, k). \quad (10)$$

We say that  $\Lambda_d$  covers  $\Lambda$  with a radius  $\rho_d$ .

This radius characterizes the density of  $\Lambda_d$  inside  $\Lambda$ . Given any  $\lambda$  in  $\Lambda$ , one is guaranteed that there exists an element  $k$  of  $\Lambda_d$  close to  $\lambda$ , i.e. within a geodesic distance  $\rho_d$ .

**Theorem 1.** *Given a Hilbert space  $\mathcal{S} \subseteq L^2(X)$  with a non zero greedy factor  $\beta$ , and a dictionary  $\mathcal{D} = \text{dict}(\Lambda) \subset \mathcal{S}$  of non-zero condition number  $\mathcal{K}^{-1}$ , if  $\Lambda_d$  covers  $\Lambda$  with radius  $\rho_d$ , and if  $\rho_d < \beta/\sqrt{1+\mathcal{K}}$ , then, for functions belonging to  $\mathcal{S}$ , a  $\text{dMP}(\alpha)$  algorithm using  $\mathcal{D}_d = \text{dict}(\Lambda_d)$  is bounded by the exponential convergence rate of a  $\text{cMP}(\alpha')$  using  $\mathcal{D}$  with a weakness parameter given by  $\alpha' = \alpha(1 - \beta^{-2} \rho_d^2 (1 + \mathcal{K}))^{1/2} < \alpha$ .*

*Proof.* Notice first that since  $f \in \mathcal{S}$  and  $\mathcal{D}_d \subset \mathcal{D} \subset \mathcal{S}$ ,  $R^m f \in \mathcal{S}$  for all iteration  $m$  of  $\text{dMP}$ . Let us take the  $(m+1)^{\text{th}}$  step of  $\text{dMP}(\alpha)$  and write  $u = R^m f$ . We have of course  $\|R^{m+1} f\|^2 = \|u\|^2 - S_u(k_{m+1})$ , where  $k_{m+1}$  is the atom obtained from the selection rule (3c), i.e.  $S_u(k_{m+1}) \geq \alpha^2 \sup_{k \in \Lambda_d} S_u(k)$ .

Denote by  $g_{\tilde{\lambda}}$  the atom of  $\mathcal{D}$  that best represents  $R^m f$ , i.e.  $S_u(\tilde{\lambda}) = \sup_{\lambda \in \Lambda} S_u(\lambda)$ . If  $\tilde{k}$  is the closest element of  $\tilde{\lambda}$  in  $\Lambda_d$ , we have  $d_{\mathcal{G}}(\tilde{\lambda}, \tilde{k}) \leq \rho_d$  from the covering property of  $\Lambda_d$ , and the Proposition 1 tells us that, with  $u = R^m f$ ,  $|S_u(\tilde{k}) - S_u(\tilde{\lambda})| \leq \rho_d^2 (1 + \mathcal{K}) \|u\|^2$ , since  $\partial_i S_u(\tilde{\lambda}) = 0$  for all  $i$ .

Therefore,  $S_u(\tilde{k}) \geq S_u(\tilde{\lambda}) - \rho_d^2 (1 + \mathcal{K}) \|u\|^2 \geq \beta^2 \|u\|^2 - \rho_d^2 (1 + \mathcal{K}) \|u\|^2$ , and  $S_u(\tilde{k}) \geq \beta^2 (1 - \beta^{-2} \rho_d^2 (1 + \mathcal{K})) \|R^m f\|^2$ , this last quantity being positive from the density requirement, i.e.  $\rho_d < \beta/\sqrt{1+\mathcal{K}}$ .

In consequence,  $S_u(k_{m+1}) \geq \alpha^2 \sup_{k \in \Lambda_d} S_u(k) \geq \alpha^2 S_u(\tilde{k})$ , implying  $\|R^{m+1} f\|^2 = \|u\|^2 - S_u(k_{m+1}) \leq \|u\|^2 - \alpha^2 S_u(\tilde{k}) \leq \|u\|^2 (1 - \alpha'^2 \beta^2)$ , for  $\alpha' \triangleq \alpha(1 - \beta^{-2} \rho_d^2 (1 + \mathcal{K}))^{1/2}$ . So,  $\|R^{m+1} f\| \leq (1 - \alpha'^2 \beta^2)^{(m+1)/2} \|f\|$ , which is the exponential convergence rate of the Weak( $\alpha$ ) Matching Pursuit in  $\mathcal{D}$  when  $\beta$  exists [29, 30].  $\square$

The previous proposition has an interesting interpretation : a weak Matching Pursuit decomposition in a discrete dictionary corresponds, in terms of rate of convergence, to a weaker Matching Pursuit in the continuous dictionary from which the discrete one is extracted.



About the hypotheses of the proposition, notice first that the existence of a greedy factor inside  $\mathcal{S}$  concerns the continuous dictionary  $\mathcal{D}$  and not the discrete one  $\mathcal{D}_d$ . Consequently, this condition is certainly easier to fulfill from the high redundancy of  $\mathcal{D}$ . Second, the *density requirement*,  $\rho_d < \beta/\sqrt{1+\mathcal{K}}$ , is just sufficient since the Proposition 1 does not state that it achieves the best bound for the control of  $|S_u(\lambda_a) - S_u(\lambda_b)|$  when  $\lambda_a$  is critical. It is interesting to note that this inequality relates  $\rho_d$ , a quantity that characterizes the discretization  $\Lambda_d$  in  $\Lambda$ , to  $\beta$  and  $\mathcal{K}$ , which depend only on the dictionary. In particular,  $\beta$  represents the density of  $\mathcal{D}$  inside  $\mathcal{S} \subset L^2(X)$ , and  $\mathcal{K}$  depends on the shape of the atoms through the curvature of the dictionary.

Finally note that as  $\beta < 1$  (from definition (4)) and  $\mathcal{K} > 1$  (Prop. 2), the density radius must at least satisfy  $\rho_d < \frac{1}{\sqrt{2}}$  to guarantee that our analysis is valid.

## 4.2 A Simple Example of Discretization

Let us work on the *line* with  $L^2(X) = L^2(\mathbb{R}, dt)$ , and check if the hypothesis of the previous theorem can be assessed in the simple case of an *affine* (wavelet-like) dictionary.

We select a symmetric and real mother function  $g \in L^2(\mathbb{R})$  well localized around the origin, e.g. a Gaussian or a Mexican Hat, normalized such that  $\|g\| = 1$ . The parameter set  $\Lambda$  is related to the *affine group*, the group of translations and dilations  $G_{\text{aff}}$ . We identify  $\lambda = (\lambda^0 = b, \lambda^1 = a)$ , where  $b \in \mathbb{R}$  and  $a > 0$  are the translation and the dilation parameters respectively. The dictionary  $\mathcal{D}$  is defined from the atoms  $g_\lambda(t) = [U(\lambda)g](t) = a^{-1/2} g((t-b)/a)$ , with  $\|g_\lambda\| = 1$  for all  $\lambda \in \Lambda$ . Our atoms are nothing but the wavelets of a Continuous Wavelet Transform if  $g$  is admissible [35], and  $U$  is actually the representation of the affine group on  $L^2(\mathbb{R})$  [36].

In the technical report [31], we prove that the associated metric is given by  $\mathcal{G}_{ij}(\lambda) = a^{-2} W$ , where  $W$  is a constant  $2 \times 2$  diagonal matrix depending only of the mother function  $g$  and its first and second derivatives. Since  $\mathcal{G}^{ij}(\lambda) = a^2 W^{-1}$ ,  $\mathcal{K}$  can be bounded by a constant also associated to  $g$  and its first and second order time derivatives.

Finally, given the  $\tau$ -adic parameter discretization

$$\Lambda_d = \{k_{jn} = (b_{jn}, a_j) = (nb_0\tau^j, a_0\tau^j) : j, n \in \mathbb{Z}\},$$

with  $\tau > 1$  and  $a_0, b_0 > 0$ , the density radius  $\rho_d$  of  $\Lambda_d$  is shown to be bounded by  $\rho_d \leq Ca_0^{-1}b_0 + D \ln \tau$ , with  $C$  and  $D$  depending only of the norms of  $g$  and its first derivative.

This bound has two interesting properties. First, as for the grid  $\Lambda_d$ , it is invariant under the change  $(b_0, a_0) \rightarrow (2b_0, 2a_0)$ . Second, it is multiplied by  $2^n$  if we realize a “zoom” of factor  $2^n$  in our  $\tau$ -adic grid, in other words, if  $(b_0, \tau) \rightarrow (2^n b_0, \tau^{2^n})$ . By the same argument, the true density radius has also to respect these rules. Therefore, we conjecture that  $\rho_d = C'a_0^{-1}b_0 + D' \ln \tau$ , for two particular (non computed) positive constants  $C'$  and  $D'$ .

Unfortunately, even for this simple affine dictionary, the existence of  $\beta = \beta(\mathcal{S}, \mathcal{D})$  is non trivial to prove. However, if the greedy factor exists, the control of  $\tau$ ,  $a_0$  and  $b_0$  over  $\rho_d$  tells us that it is possible to satisfy the density requirement for convenient values of these parameters.

## 5 Optimization of Discrete Matching Pursuits

The previous section has shown that under a few assumptions a dMP is equivalent, in terms of rate of convergence, to a weaker cMP in the continuous dictionary from which the discrete one

has been sampled.

**Question 2.** *Can we improve the rate of convergence of a dMP, not with an obvious increasing of the dictionary sampling, but by taking advantage of the dictionary geometry ?*

Our approach is to introduce an optimization of the discrete dMP scheme. In short, at each iteration, we propose to use the atoms of  $\mathcal{D}_d$  as the seeds of an iterative optimization, such as the basic *gradient descent/ascent*, respecting the geometry of the manifold  $\mathcal{M} = (\Lambda, \mathcal{G}_{ij})$ .

Under the same density hypothesis of Theorem 1, we show that in the worst case and if the number of optimization steps is large enough, an optimized discrete MP is again equivalent to a continuous dMP, but with a weakness factor two times closer to unity than for the non-optimized discrete MP.

In this section, we first introduce the basic gradient descent/ascent on a manifold. Next, we show how this optimization can be introduced in the Matching Pursuit scheme to defined the geometrically optimized MP (gMP). Finally, the rate of convergence of this method is analyzed.

### 5.1 Gradient Ascent on Riemannian Manifolds

Given a function  $u \in L^2(X)$  and  $S_u(\lambda) = \langle g_\lambda, u \rangle^2$ , we wish to find the parameter that maximizes  $S_u$ , i.e.

$$\lambda_* = \arg \max_{\lambda \in \Lambda} S_u(\lambda) \quad (\text{P.1})$$

Equivalently, by introducing  $h_{u,\lambda} = \langle g_\lambda, u \rangle g_\lambda$ , we can decide to find  $\lambda_*$  by the minimization

$$\lambda_* = \arg \min_{\lambda \in \Lambda} \|u - h_{u,\lambda}\|^2. \quad (\text{P.2})$$

If we are not afraid to get stuck on local maxima (P.1) or minima (P.2) of these two non-convex problems, we can solve them by using well known optimization techniques such as gradient descent/ascent, or Newton or Newton-Gauss optimizations.

We present here a basic gradient ascent of the Problem (P.1) that respect the geometry of  $\mathcal{M} = (\Lambda, \mathcal{G}_{ij})$  [37]. This method increases iteratively the value of  $S_u$  by following a path in  $\Lambda$ , composed of geodesic segments, driven by the gradient of  $S_u$ .

Given a sequence of step size  $t_r > 0$ , the gradient ascent of  $S_u$  starting from  $\lambda_0 \in \Lambda$  is defined by the following induction [38] :

$$\phi_0(\lambda_0) = \lambda, \quad \phi_{r+1}(\lambda_0) = \gamma(t_r, \phi_r(\lambda_0), \xi_r(\lambda_0)),$$

where  $\xi_r(\lambda_0) = |\nabla S_u(\phi_r(\lambda_0))|^{-1} \nabla S_u(\phi_r(\lambda_0))$  is the *gradient direction* obtained from the gradient  $\nabla^i S_u = \mathcal{G}^{ij} \partial_j S_u$ , and  $\gamma(s, \lambda_0, \xi_0)$  is the geodesic starting at  $\lambda_0 = \gamma(0, \lambda_0, \xi_0)$  with the unit velocity  $\xi_0 = \frac{\partial}{\partial s} \gamma(0, \lambda_0, \xi_0)$ . Notice that  $\nabla^i$  is the natural notion of gradient on a Riemannian manifold. Indeed, as for the Euclidean case, with  $\nabla^i h \triangleq \mathcal{G}^{ij} \partial_j h$  for  $h \in L^2(X)$ , given  $w \in T_\lambda \Lambda$ , the directional derivative  $D_w h$  is equivalent to  $D_w h(\lambda) \triangleq w^i \partial_i h(\lambda) = \langle \nabla h, w \rangle_\lambda \triangleq w^i \nabla^j h(\lambda) \mathcal{G}_{ij}(\lambda)$ , since  $\mathcal{G}^{ik} \mathcal{G}_{kj} = \delta_j^i$ .

Practically, in our gradient ascent, we use the linear first order approximation of  $\gamma$ , i.e.

$$\phi_{r+1}(\lambda) = \phi_r(\lambda) + t_r \xi_r(\lambda), \quad (11)$$

valid for small value of  $t_r$  (error in  $O(t_r^2)$ ). This is actually an optimization method since  $\partial_i S_u(\phi_r(\lambda)) \xi_r^i = |\partial S_u(\phi_r(\lambda))| > 0$  and  $S_u(\phi_{r+1}(\lambda)) = S_u(\phi_r(\lambda)) + t_r |\partial S_u(\phi_r(\lambda))| + O(t_r^2) \geq S_u(\phi_r(\lambda))$ , for a convenient step size  $t_r > 0$ . At each step of this gradient ascent, the value  $t_r$  is chosen so that  $S_u$  is increased. This can be done for instance by a *line search* algorithm [39]. From the positive definiteness of  $\mathcal{G}_{ij}$  and  $\mathcal{G}^{ij}$ , a fixed point  $\phi_{r+1}(\lambda) = \phi_r(\lambda)$  is reached if  $\nabla^i S_u(\phi_r(\lambda)) = \partial_i S_u(\phi_r(\lambda)) = 0$  for all  $i$ .

More sophisticated algorithms such as Newton or Newton-Gauss can be developed to solve the Problem (P.2) on a Riemannian manifolds [38, 40] even if, unlike to the flat case, a direct definition of the Hessian does not exist on differentiable manifolds. However, we will not use them here as our aim is to prove that a dMP driven by the very basic optimization above provides already a better rate of convergence than the non-optimized dMP.

## 5.2 Optimized Discrete Matching Pursuit Algorithm

Let us optimize each step of a discrete MP using the gradient ascent of the previous section.

**Definition** Given sequence of positive integers  $\kappa_m$  and a weakness factor  $0 < \alpha \leq 1$ , the geometrically optimized discrete matching pursuit (gMP( $\alpha$ )) is defined by

$$R^0 f = f \quad (\text{initialization}), \quad (12a)$$

$$R^{m+1} f = R^m f - \langle g_{\nu_{m+1}}, R^m f \rangle g_{\nu_{m+1}}, \quad (12b)$$

$$\langle g_{\nu_{m+1}}, R^m f \rangle^2 \geq \alpha^2 \sup_{k \in \Lambda_d} \langle g_{\phi_{\kappa_m}(k)}, R^m f \rangle^2. \quad (12c)$$

Notice that the best atom  $g_{\nu_{m+1}}$  is selected in the set  $\Phi_m \triangleq \{g_{\phi_{\kappa_m}(k)} : k \in \Lambda_d\} \subset \mathcal{D}$ . Elements of  $\Phi_m$  are determined by applying the *optimization function*  $\phi_r : \Lambda_d \rightarrow \Lambda$  of our gradient ascent defined in (11) on elements of  $\Lambda_d$ . In consequence,  $\Phi_m$  depends on  $R^m f$  and is thus different at each iteration  $m$ .

**Rate of convergence** The following theorem characterizes the rate of convergence of the optimized Matching Pursuit defined in (12).

**Theorem 2.** *Given the notations and the conditions of Theorem 1, there exists a sequence of positive integers  $\kappa_m$  such that, the gMP( $\alpha$ ) decomposition of functions in  $\mathcal{S} \subset L^2(X)$  optimized  $\kappa_m$  steps at each iteration  $m$ , is bounded by the same rate of convergence as a cMP( $\alpha''$ ) using the corresponding continuous dictionary  $\mathcal{D}$  with  $\alpha'' = \alpha(1 - \frac{1}{2} \beta^{-2} \rho_d (1 + \mathcal{K}))^{1/2} \leq \alpha$ .*

In other words, for  $\alpha = 1$ , a gMP is equivalent to a cMP with a weakness factor two times closer to unity than the one reached by a dMP in the same conditions. Before proving this result, let us introduce some new lemmata.

**Lemma 1.** *Given a function  $u \in L^2(X)$  and a dictionary  $\mathcal{D}$  of non-zero condition number  $\mathcal{K}^{-1}$ , if  $\lambda_a$  is critical in the direction of  $\lambda_b$ , and if  $\lambda_b$  is critical in the direction of  $\lambda_a$ , i.e.  $\gamma'^i(0) \partial_i S_u(\lambda_a) = \gamma'^i(d) \partial_i S_u(\lambda_b) = 0$  for  $\gamma = \gamma_{\lambda_a \lambda_b}$  the geodesic joining  $\lambda_a$  and  $\lambda_b$  and  $d = d_{\mathcal{G}}(\lambda_a, \lambda_b)$ , then*

$$|S_u(\lambda_a) - S_u(\lambda_b)| \leq \frac{1}{2} \|u\|^2 d_{\mathcal{G}}(\lambda_a, \lambda_b)^2 (1 + \mathcal{K}). \quad (13)$$

*Proof.* Without loss of generality, assume that  $S_u(\lambda_a) \geq S_u(\lambda_b)$ . If this is not the case, we can switch the labels  $a$  and  $b$ . Let us define  $\lambda(\theta) = \gamma(\theta d)$  with  $\theta \in [0, 1]$  on the geodesic  $\gamma = \gamma_{\lambda_a \lambda_b}$ . We have  $\lambda_a = \lambda(0)$  and  $\lambda_b = \lambda(1)$ . Using the Corollary 1, the two following inequalities hold :  $S_u(\lambda(\theta)) \geq S_u(\lambda_a) - \|u\|^2 d_{\mathcal{G}}(\lambda(\theta), \lambda_a)^2 (1 + \mathcal{K})$  and  $S_u(\lambda(\theta)) \leq S_u(\lambda_b) + \|u\|^2 d_{\mathcal{G}}(\lambda(\theta), \lambda_b)^2 (1 + \mathcal{K})$ .

Therefore, since by definition of  $\lambda(\theta)$ ,  $d_{\mathcal{G}}(\lambda(\theta), \lambda_a) = \theta d$  and  $d_{\mathcal{G}}(\lambda(\theta), \lambda_b) = (1 - \theta)d$ , we find  $S_u(\lambda_a) - S_u(\lambda_b) \leq \|u\|^2 (\theta^2 + (\theta - 1)^2) d^2 (1 + \mathcal{K})$  for all  $\theta \in [0, 1]$ . Taking the minimum over all  $\theta$ , we obtain finally  $S_u(\lambda_a) - S_u(\lambda_b) \leq \frac{1}{2} \|u\|^2 d_{\mathcal{G}}(\lambda_a, \lambda_b)^2 (1 + \mathcal{K})$ .  $\square$

In other words, the critical nature of  $\lambda_a$  and  $\lambda_b$  divides by two the bound on the decreasing of  $S_u$  between them compared to the situation where only one of these points is critical.

**Lemma 2.** *Given a function  $u \in L^2(X)$ , assume that  $S_u(\lambda)$  has a global maximum at  $\lambda_M$ , i.e.  $\partial_i S_u(\lambda_M) = 0$  for all  $i$ , and write  $\mathcal{T}_k = \{\phi_r(k) : r \in \mathbb{N}\}$  the trajectory of the gradient ascent described in (11) starting from a point  $k \in \Lambda_d$ . There exists a  $\lambda' \in \mathcal{T}_k$  that can be reached in a finite number of optimization steps, such that*

$$S_u(\lambda_M) - S_u(\lambda') \leq \frac{1}{2} \|u\|^2 d_{\mathcal{G}}(\lambda_M, k)^2 (1 + \mathcal{K}). \quad (14)$$

For the sake of clarity, the proof of this technical Lemma is placed in Appendix C. The main idea is to find a point in the trajectory  $\mathcal{T}_k$  that is closer to  $\lambda_M$  than  $k$ , and that is also critical in the direction of  $\lambda_M$  so that Lemma 1 can be applied. Let us now enter in the proof of the previous proposition.

*Proof of Theorem 2.* In our gMP( $\alpha$ ) decomposition of a function  $f \in \mathcal{S} \subset L^2(X)$  defined before, given the iteration  $m + 1$  where  $u = R^m f$  is analyzed, denote by  $\tilde{\lambda}$  the parameter of the atom in  $\mathcal{D}$  maximizing  $S_u$ , i.e.  $S_u(\tilde{\lambda}) = \sup_{\lambda \in \Lambda} S_u(\lambda)$ .

If  $\tilde{k}$  is the closest element of  $\Lambda_d$  to  $\tilde{\lambda}$ , from the covering property of  $\Lambda_d$  we have  $d_{\mathcal{G}}(\tilde{\lambda}, \tilde{k}) \leq \rho_d$ , and the Lemma 2 tells us that there exists a finite number of optimization steps  $\kappa_m$  such that  $S_u(\phi_{\kappa_m}(\tilde{k})) \geq S_u(\tilde{\lambda}) - \frac{1}{2} \rho_d \|u\|^2 (1 + \mathcal{K}) \geq \beta^2 (1 - \frac{1}{2} \beta^{-2} \rho_d^2 (1 + \mathcal{K})) \|u\|^2$ , where the last term is positive from the density requirement  $\rho_d < \beta / \sqrt{1 + \mathcal{K}}$ .

Therefore, from the selection rule (12c),  $S_u(\nu_{m+1}) \geq \alpha^2 S_u(\phi_{\kappa_m}(\tilde{k}))$ . We have thus  $\|R^{m+1} f\|^2 = \|u\|^2 - S_u(\nu_{m+1}) \leq \|u\|^2 - \alpha^2 S_u(\phi_{\kappa_m}(\tilde{k})) \leq \|u\|^2 (1 - \alpha'' \beta^2)$ , with  $\alpha'' \triangleq \alpha (1 - \frac{1}{2} \beta^{-2} \rho_d^2 (1 + \mathcal{K}))^{1/2}$ . So,  $\|R^{m+1} f\| \leq (1 - \alpha'' \beta^2)^{(m+1)/2} \|f\|$  which is also the exponential convergence rate of the cMP( $\alpha''$ ) in  $\mathcal{D}$  when  $\beta$  exists.  $\square$

In Theorem 2, even if the sequence of optimization steps  $\kappa_m$  is proved to exist, it is actually unknown. One practical way to overcome this problem is to observe how the ratio  $\frac{|\nabla S_u|}{S_u}$  decreases at each optimization steps, and to stop the procedure once this value falls below a predefined threshold. This follows from the idea that the closer to a local maximum  $S_u(\phi_r(k))$  is, the smaller must be the optimization step. As it is often the case in optimization problems, an upper bound on the number of optimization steps can be fixed jointly to this threshold test.

## 6 Experiments

In this section, dMP and gMP decompositions of 1-D and 2-D signals are studied experimentally in different situations. These will imply different classes of signals and different discretization of parametrization of various densities.

Prior to these experiments, some remarks have to be made about dMP and gMP implementations. First, for both algorithms, as described in Equations (3) and (12), a *full-search* has to be performed in  $\mathcal{D}_d = \text{dict}(\Lambda_d)$  to compute all the squared scalar products  $S_u$  of the current residue  $u = R^m f$ , with atoms  $g_k$ . We decide thus to reduce the computational complexity of this full-search with the help of the Fast Fourier Transform (FFT). One component (for 1-D signals) or two components (for 2-D signals) of the parametrization correspond indeed to a regular grid of atoms positions, which makes  $S_u$  a discrete correlation relatively to these parameters. Moreover, as described in detail in [41,42], we apply the fast *boundary renormalization* of atoms, where atoms of  $\mathcal{D}$  truncated by the limit of the signal remain valid atoms, i.e. of unit norm, and features that suddenly terminate at the signal boundary are correctly caught in the procedure. Notice that all our dMP and gMP experiments are performed in the non-weak case, i.e.  $\alpha = 1$ .

Second, for the Gradient-Ascent optimization, we realize some simplifications to the initial formulation : the best discrete atom only is optimized at each MP iteration and  $\kappa_m = \kappa > 0$  for all  $m$ , with  $\kappa$  typically equal to 5 or 10. Even if these two restrictions are not optimal compared to the method described in the theoretical results, the gain of the optimization in the quality of signals reconstructions is already impressive. We also set all the step sizes to  $t_r = \chi > 0$ , with  $\chi = 0.1$  in all our experiments. Then, at each optimization step  $r$ , we adaptively decrease the step parameter  $t_r$  by dividing it by 2 if the ascent condition is not met, i.e. if  $S_u(\phi_{r+1}(k)) < S_u(\phi_r(k))$ . If after 10 divisions, the ascent condition still does not hold, the optimization process is simply stopped.

Finally, let us mention that our algorithms are written in MATLAB© and are consequently not truly optimized. The different computation times that we provide through this section allow us only to compare various schemes, as for dMP and gMP decomposition of the same signal. All of our experiments were realized on a Pentium 1.73 GHz with 1Gb of memory.

## 6.1 One Dimensional Analysis

This section analyzes the benefit obtained from gMP, and from an increased density of the discrete dictionary, when decomposing some specific classes of randomly generated 1-D signals. In our experiments, each 1-D signal is of unit norm and has  $N = 2^{13}$  samples. Each signal consists of the sum of 100 random bursts, each burst being a rectangular or Gaussian window, depending on the class of the signal. The position and magnitude of each burst is selected at random, according to a uniform distribution. The duration of the rectangular window and the standard deviation of the Gaussian function are selected uniformly within the range  $[\frac{1}{2}L, \frac{3}{4}L]$ , for  $L = 2^8$ . The mother function of the dictionary is the *Mexican Hat* function  $g(t) \propto (1 - t^2) e^{-t^2/2}$ . Its scale and translation parameters are sampled as defined in Section 4.2, following the  $\tau$ -adic discretization  $\Lambda_d = \{(nb_0\tau^j, a_0\tau^j) : j, n \in \mathbb{Z}\}$ , with  $a_0 = 1$ . We work in the non-weak case, i.e.  $\alpha = 1$ , for dMP and gMP, and we set  $\kappa = 10$  for gMP.

Figures 1(a) and 1(b) analyze how the energy  $\|R^m f\|^2$  of the residual decreases with the number  $m$  of MP iterations for the random Gaussian and rectangular signals, respectively. Notice that only a small number of iterations are studied (twelve) since our analysis aims at analyzing the behaviour of dMP and gMP on one class of signals. However the current residual  $R^m f$  belongs only approximately to the considered class on small  $m$  when not many atoms have been subtracted to  $f = R^0 f$ . Results presented are averaged over 20 trials. In each graph, two

distinct discretizations of the Mexican Hat parameters are considered to provide two discrete dictionaries, with one ( $b_0 = 1, \log_2 \tau = 0.25$ ) being two times denser than the other ( $b_0 = 2, \log_2 \tau = 0.5$ ), according to the behavior<sup>11</sup> of the density radius  $\rho_d$  analyzed in Section 4.2. Both discrete and geometrically optimized MP are studied for each dictionary. We observe that gMP significantly outperforms dMP, and that an increased density of the dictionary also speeds up the MP convergence. By comparing Figure 1(a) and 1(b), we also observe that the residual energy decreases much faster for Gaussian signals than for rectangular ones, which unsurprisingly reveals that the Mexican Hat dictionary is better suited to represent Gaussian structures.

Figures 1(c)-1(f) further analyze the impact of the discretization of the dictionary parameters on MP convergence. In these figures, we introduce the notion of *normalized atom energy* (NAE) to measure the convergence rate of a particular dictionary dealing with a specific class of signals at a specific MP iteration step. Formally, the NAE denotes the expected value of the best squared atom coefficient computed on a normalized signal when this one is randomly generated within a specific class of signals. Mathematically,  $\text{NAE} = \mathbb{E}[\langle g_{\lambda_*}, \frac{u}{\|u\|} \rangle^2]$ , where  $u$  is a sample signal of the class and the  $g_{\lambda_*}$  the associated best atom for a fixed greedy algorithm (dMP or gMP). We show the dependence of NAE on the discretization for the 1<sup>st</sup> and 30<sup>th</sup> iteration<sup>12</sup> for both rectangular and Gaussian signals. Results are averaged over 500 trials.

By considering the dMP and gMP curves in Figures 1(c)-1(f), we first observe that the NAE is significantly higher for gMP than for dMP, which confirms the advantage of using gradient ascent optimization to refine the parameters of the atoms extracted by dMP. Note that the NAE for a Gaussian random signal (Fig. 1(c)-1(d)) is nearly one order of magnitude higher than for a rectangular one (Fig. 1(e)-1(f)). This confirms that the Mexican Hat dictionary better matches the Gaussian structures than the rectangular ones. In all cases, the NAE sharply decreases with the iteration index, which is not a surprise as the coherence between the signal and the dictionary decreases as MP expansion progresses.

To better understand the penalty induced by the discretization of the continuous dictionary, we now analyze how the rate of convergence for a particular class of signals behaves compared to the reference provided by a signal composed of a single Mexican Hat function. For that purpose, an additional curve, denoted  $\text{dMP}_a$ , has been plotted in each graph. This curve is expected to provide an upper bound to the penalty induced by a sparser dictionary. Specifically,  $\text{dMP}_a$  plots the energy captured during the 1<sup>st</sup> step of the dMP expansion of a random (scale and position) Mexican Hat function, as a function of the discretization parameter  $\log_2 \tau$ . As the Mexican Hat is the generative function of the dictionary, the 1<sup>st</sup> step of the MP expansion would capture the entire function energy if the entire continuous dictionary were used, but is particularly penalized by a discretization of the dictionary. In each graph of Figures 1(c)-1(f), to compare  $\text{dMP}_a$  with dMP, the  $\text{dMP}_a$  curve obtained with pure atoms (i.e. unit coefficients) is scaled to correspond to atoms whose energy is set to the NAE expected from the expansion of the corresponding class of signals with a continuous dictionary. In practice, the NAE expected with a continuous dictionary is estimated based on the NAE computed with gMP and the densest dictionary ( $\log_2 \tau = 0.25$ ).

<sup>11</sup>Obviously equivalent for  $\log_2 \tau$  or  $\ln \tau$  variations.

<sup>12</sup>Note that the NAE at the 30<sup>th</sup> iteration refers to the NAE computed on the residual signals obtained after 29 iterations of the gMP with the densest dictionary, independently of the actual discrete dictionary considered at iteration 30. Hence, the reference class of signals to compute the NAE at iteration 30 is the same for all investigated dictionaries, i.e. for all  $\log_2 \tau$  values.



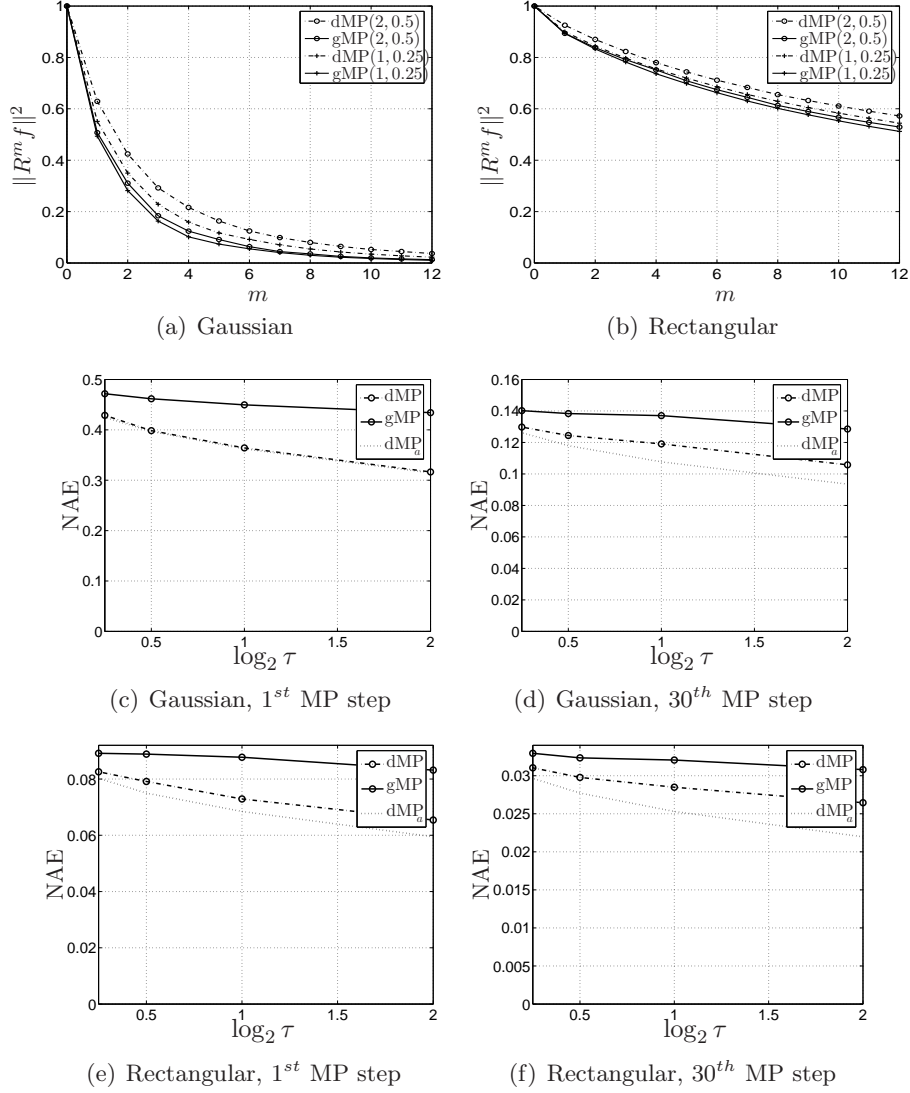


Figure 1: (a)-(b) Residual energy as a function of the MP iteration.  $\text{dMP}(b_0, \log_2 \tau)$  and  $\text{gMP}(b_0, \log_2 \tau)$  refer to discrete and optimized MP, computed on a discretization  $\Lambda_d = \{(nb_0\tau^j, a_0\tau^j) : j, n \in \mathbb{Z}\}$  of the continuous Mexican Hat dictionary. (c)-(f) Normalized atom energy (NAE) as a function of the  $\log_2 \tau$  discretization parameter.  $b_0$  is set to one in all cases. dMP and gMP respectively refer to discrete and optimized MP. dMP<sub>a</sub> provides a lower bound to the decrease of NAE with  $\log_2 \tau$ , and is formally described in the text.

The approximation is reasonable as we observe that gMP saturates for small  $\log_2 \tau$  values, i.e. for large densities. We first observe that both the dMP and the dMP<sub>a</sub> curves nearly coincide in Figure 1(c). Hence, the MP expansion of a Gaussian signal is penalized as much as the one of a Mexican Hat function by a reduction of the discrete dictionary density. We then observe that the penalty induced by a reduction of density decreases as the coherence between signal and dictionary structures drops. This is for example the case when the signal to represent is intrinsically sharper than the dictionary structures (Fig. 1(e)-1(f)), or because the coherent structures have been extracted during the initial MP steps (Fig. 1(d)). This last observation is of practical importance because it reveals that using a coarsely discretized dictionary incurs a greater penalty during the first few iterations of the MP expansion than during the subsequent ones. For compression applications, it might thus be advantageous to progressively decrease the density of the dictionary along the expansion process, the cost associated to the definition of the atom indices decreasing with the density of the dictionary<sup>13</sup>. Hence, it might be more efficient – in a rate-distortion sense – to use a dense but expensive dictionary during the first MP iterations, so as to avoid penalizing the MP convergence rate, but a sparser and cheaper during subsequent steps, so as to save bits. We plan to investigate this question in details in a future publication.

## 6.2 Two Dimensional Analysis

This section analyzes experimentally the effect of discretizing a dictionary on the Matching Pursuit decomposition of images, i.e. with the Hilbert space  $L^2(\mathbb{R}^2)$ .

**Parametrization and Dictionary** We use the same dictionary as in [41]. Its mother function  $g$  is defined by a separable product of two 1-D behaviors : a Mexican Hat wavelet in the  $x$ -direction, and a Gaussian in the  $y$ -direction, i.e.  $g(\mathbf{x}) = (\frac{4}{3\pi})^{1/2} (1 - x^2) \exp(-\frac{1}{2} |\mathbf{x}|^2)$ , where  $\mathbf{x} = (x, y) \in \mathbb{R}^2$  and  $\|g\| = 1$  [30]. Notice that  $g$  is infinitely differentiable.

The dictionary is defined by the translations, rotations, and anisotropic dilations of  $g$ . Mathematically, these transformations are represented by operators  $T_{\mathbf{b}}$ ,  $R_{\theta}$ , and  $D_{\mathbf{a}}$ , respectively. These are given by  $[T_{\mathbf{b}} g](\mathbf{x}) = g(\mathbf{x} - \mathbf{b})$ ,  $[R_{\theta} g](\mathbf{x}) = g(r_{\theta}^{-1} \mathbf{x})$ , and  $[D_{\mathbf{a}} g](\mathbf{x}) = (a_1 a_2)^{-1/2} g(d_{\mathbf{a}}^{-1} \mathbf{x})$ , for  $\theta \in S^1 \simeq [0, 2\pi)$ ,  $\mathbf{b} \in \mathbb{R}^2$ ,  $\mathbf{a} = (a_1, a_2)$ ,  $a_1, a_2 \in \mathbb{R}_+^*$ , while  $r_{\theta}$  is the usual  $2 \times 2$  rotation matrix  $r_{\theta}$  and  $d_{\mathbf{a}} = \text{diag}(a_1, a_2)$ .

In other words, we have a parametrization of  $P = 5$  dimensions and  $\Lambda = \{\lambda = (\lambda^0, \dots, \lambda^4) = (b_1, b_2, \theta, a_1, a_2) \in \mathbb{R}^2 \times S^1 \times (\mathbb{R}_+^*)^2\}$ . At the end, each atom of the dictionary  $\mathcal{D} = \{g_{\lambda} : \lambda \in \Lambda\}$  is generated by  $g_{\lambda}(\mathbf{x}) = [U(\lambda) g](\mathbf{x}) \triangleq [T_{\mathbf{b}} R_{\theta} D_{\mathbf{a}} g](\mathbf{x})$ , with  $\|g_{\lambda}\| = \|g\| = 1$ .

Obviously, the dictionary  $\mathcal{D}$  is complete in  $L^2(\mathbb{R}^2)$ . Indeed, translations, rotations and isotropic dilations alone are already enough to constitute a wavelet basis of  $L^2(X)$  since  $g$  is an admissible wavelet [35, 43]. Finally, as requested in the previous section, from the smoothness of  $g$  and of the transformations  $U$  above, the atoms  $g_{\lambda}$  of our dictionary  $\mathcal{D}$  are twice differentiable on each component  $\lambda^i$ .

**Spatial Sampling** For all our experiments, images are discretized on a Cartesian regular grid of pixels, i.e. an image  $f$  takes its values on the grid  $\mathcal{X} = ([0, N_x] \times [0, N_y]) \cap \mathbb{Z}^2$ , with  $N_x$  and  $N_y$  the “ $x$ ” and “ $y$ ” sizes of the grid. We work in the *continuous approximation*, that is we assume

<sup>13</sup>Less distinct atom indices need to be described by the codewords.



Figure 2: 300 atoms reconstruction results. (a) dMP :  $J = 5$ ,  $K = 8$ , PSNR: 26.63 dB, 4634s. (b) gMP :  $J = 3$ ,  $K = 4$ ,  $\kappa = 10$ , PSNR: 26.68 dB, 949s.

	$J = 3$	$J = 5$
$K = 4$	24.30 dB (834s)	25.88 dB (2327s)
$(\kappa = 5)$	26.08 dB (889s)	27.09 dB (2381s)
$(\kappa = 10)$	26.68 dB (950s)	27.37 dB (2447s)
$K = 8$	25.21 dB (1660s)	26.63 dB (4634s)
$(\kappa = 5)$	27.05 dB (1715s)	27.92 dB (4703s)
$(\kappa = 10)$	27.44 dB (1772s)	28.09 dB (5131s)

Table 1: dMP and gMP applied on **Barbara** image. Quality (in PSNR) of the reconstruction after 300 iterations for various  $J$ ,  $K$  and  $\kappa$ . In each table cell, the first row correspond to dMP result, the second and the third rows to gMP.

that the grid  $\mathcal{X}$  is fine enough to guarantee that the scalar products  $\langle \cdot, \cdot \rangle$  and norms  $\|\cdot\|$  are well estimated from their discrete counterparts. This holds of course for band-limited functions on  $L(\mathbb{R}^2)$ .

In consequence, in order to respect this continuous approximations and to have dictionary atoms smaller than the image size, the mother function  $g$  of our dictionary  $\mathcal{D}$  must be dilated in a particular range of scales so that  $g_\lambda$  is essentially band-limited, i.e.  $a_1, a_2 \in [a_m, a_M]$ . According to the definition of  $g$  above, we set experimentally  $a_m = 0.7$  and  $a_M = \min(N_x, N_y)$ .

**Discrete Parameter Space** We decide to sample regularly  $\Lambda$  so that to have  $N_{\text{pix}} = N_x N_y$  positions  $\mathbf{b}$ ,  $J^2$  scales  $a_1$  and  $a_2$  selected logarithmically in the range  $[a_m, a_M]$ , and  $K$  orientations evenly spaced in  $[0, \pi)$ , with  $J, K \in \mathbb{N}$ . At the end, we obtain the discretized parameter set  $\Lambda_d = \Lambda_d(N_{\text{pix}}, J, K) = \{(\mathbf{b}, \theta_n, a_{1j}, a_{2j'}), \mathbf{b} \in \mathcal{X}, n \in [0, K-1], j, j' \in [0, J-1]\}$ , and the corresponding dictionary  $\mathcal{D}_d(N_{\text{pix}}, J, K) = \text{dict}(\Lambda_d(N_{\text{pix}}, J, K))$ . The number of atoms in the dictionary is simply  $|\mathcal{D}_d| = J^2 K N_{\text{pix}}$ .

Image name	dMP	gMP ( $\kappa = 10$ )
Barbara	25.94 dB (2707s)	27.86 dB (2820s)
Lena	26.50 dB (2709s)	28.53 dB (2857s)
Baboon	24.06 dB (2770s)	24.93 dB (2900s)
Cameraman	25.80 dB (2807s)	27.62 dB (2918s)
GoldHill	26.54 dB (2810s)	28.12 dB (2961s)
Peppers	24.51 dB (2853s)	26.69 dB (3013s)

Table 2: Comparison of dMP and gMP on different usual images of size  $128 \times 128$ . Computations have been performed for  $J = 4$ ,  $K = 8$ , 300 atoms. Computation times are given indicatively in parenthesis.

**Results** We start our experiment by decomposing the venerable image of **Barbara**. 300 atoms were selected by dMP and gMP for various  $J$  and  $K$ . Results are presented in Table 1. In these tests, the best quality obtained for dMP corresponds obviously to the finest grid, i.e.  $J = 5$  and  $K = 8$  (26.63 dB, Fig.2(a)), with a computational time (CT) of 4634s. With 10 optimization steps ( $\kappa = 10$ ), the gMP for the coarsest parametrization ( $J = 3$  and  $K = 4$ ) is equivalent to the best dMP result with a PSNR of 26.68 dB and a CT of only 950s, i.e. almost five time faster. This is also far better than the dMP on the same grid (24.30 dB). The visual inspection of the dMP image ( $J = 5$ ,  $K = 8$ , Fig.2(a)) and the gMP image ( $J = 3$ ,  $K = 4$ ,  $\kappa = 10$ , 2(b)) is also instructive. Most of the features of the gMP results are well represented (e.g. Barbara’s mouth, eyes, nose, hair, ...). However, the regular pattern of the chair in the background of the picture, which needs a lot of similar atoms, is poorly drawn. This can be explained by the fact that this highly directional structure has to be represented by a lot of similarly oriented and scaled atoms with similar amplitude. The fine grid of dMP has therefore more chance to correctly fit these atoms, while the gMP on its coarse grid is deviated in its optimization process to more prominent structure with higher amplitudes. Notice finally, the best optimized result (PSNR 28.09 dB) is obtained for  $\kappa = 10$  on the grid associated to  $J = 5$  and  $K = 8$  orientations.

For our second experiment, we compare dMP and gMP ( $\kappa = 10$ ) 300 atoms approximation of well known  $128 \times 128$  pixels pictures, namely **Lena**, **Baboon**, **Cameraman**, **GoldHill**, and **Peppers**, on the same parametrization grid ( $J = 4$ ,  $K = 8$ ). For a computational time slightly higher (5%) than the dMP decomposition, we reach in all cases a significantly higher PSNR with gMP than with dMP, i.e. the dB gain is within the range [0.87, 2.03].

## 7 Related Works

A similar approach to our geometric analysis of MP atom selection rule has been proposed in [24]. In that paper, a dictionary of ( $L^2$ -normalized) wavelets is seen as a manifold associate to a Riemannian metric. However, the authors restrict their work to wavelet parametrization inherited from Lie group (such as the affine group). They also work only on the  $L^2$  (dictionary) distance between dictionary atoms and do not introduce intrinsic geodesic distance. They define a discretization of the parametrization  $\Lambda$  such that, in our notations,  $\mathcal{G}_{ij} \Delta \lambda^i \Delta \lambda^j < \epsilon$ , with  $\Delta \lambda(k)$  the local width of the cell localized on  $k \in \Lambda_d$ . There is however no analysis of the effect of this discretization on the MP rate of convergence.

In [14], the author uses a 4-dimensional Gaussian chirp dictionary to analyze 1-D signals with MP algorithm. He develops a fast procedure to find the best atom of this dictionary in the representation of the current MP residual by applying a two-step search. First, by setting the chirp rate parameter to zero, the best common Gabor atom is found with full search procedure taking advantage of the FFT algorithm. Next, a ridge theorem proves that starting from this Gabor atom, the best Gaussian chirp atom can be approximated with a controlled error. The whole method is similar to the development of our optimized matching pursuit since we start also from a discrete parametrization to find a better atom in the continuous one. However, our approach is more general since we are not restricted to a specific dictionary. We use the intrinsic geometry of any smooth dictionary manifold to perform a optimization driven by a geometric gradient ascent.

## 8 Conclusions

In this paper, we have adopted a geometrical framework to study the effect of dictionary discretization on the rate of convergence associated to MP. In a first step, we have derived an upper bound for this rate using geometrical quantities inherited from the dictionary seen as a manifold, such as the geodesic distance, the condition number of the dictionary, and the covering property of the discrete set of atoms in the continuous dictionary. We have also shown in a second step how a simple optimization of the parameters selected by the discrete dictionary, can lead theoretically and experimentally to important gain in the approximation of (general) signals.

In a future study, it could be interesting to see how our methods extend to other greedy algorithms, like the Orthogonal Matching Pursuit (OMP) [44]. However, this extension has to be performed carefully since we need to characterized the convergence of continuous OMP, as it is here for the one of MP induced by the existence of a greedy factor.

Our work paves the way for future extensions and advances in two practical fields. As explained in our 1-D experiments, a first idea could be to analyze carefully the benefit – in a rate-distortion sense – of using a dense but expensive dictionary during the first MP iterations, so as to avoid penalizing the MP convergence rate, but a sparser and cheaper dictionary during subsequent steps, so as to save bits. We plan to investigate this question in details in a future publication.

Another idea is to analyze the behaviors of gMP in the Compressive Sensing (CS) formalism, that is after random projection of the signal and atoms. Matching Pursuit is already used currently as a retrieval algorithm of CS of sparse signals [45–47]. However, recent results [48] suggests also that for manifold of bounded condition number, their geometrical structure (metric, distances) is essentially preserved after random projection of their points in a smaller space than the ambient one. If a natural definition of random projection in our continuous formalism can be formulated, a natural question is thus to check if the gradient ascent technique survives after random projection of the residual and the atoms on the same subspace. This could lead to dramatic computation time reduction, up to controlled errors that could be even attenuated by the greedy iterative procedure.

## Acknowledgements

LJ wishes to thank Prof. Richard Baraniuk and his team at Rice University (Houston, TX, USA) for the helpful discussions about general “manifolds processing” and Compressive Sensing. LJ is also very grateful to R. Baraniuk for having accepted and funded him during a short postdoctoral stay at Rice University. LJ and CDV are funded by the Belgian FRS-FNRS. We would like to thank Dr. David Kenric Hammond (LTS2/EPFL, Switzerland) for his careful proofreading and the referees for valuable comments on this paper.

## A Complements on the Geometry of $(\Lambda, \mathcal{G}_{ij})$

In this short appendix, we provide some additional information on the geometrical concepts developed in Section 2. First, as explained in that section, the parameter space  $\Lambda$  of the dictionary  $\mathcal{D} = \text{dict}(\Lambda)$  is linked to a Riemannian manifold  $\mathcal{M} = (\Lambda, \mathcal{G}_{ij})$  with a structure inherited from the dictionary  $\mathcal{D} \subset L^2(X)$ . From the geodesic definition (1) and the metric relation (2), we see that the curve  $\gamma_{\lambda_a \lambda_b}(t) \in \Lambda$  is thus also a geodesic in  $\mathcal{M}$ . In other words, it is defined only from the metric  $\mathcal{G}_{ij}$  and not anymore from the full behavior of atoms of  $\mathcal{D} \subset L^2(X)$ . In [31], we explain also that  $\mathcal{M}$  is in fact an *immersed manifold* [28] in the Hilbert manifold  $\mathcal{D} \subset L^2(X)$ , and  $\mathcal{G}_{ij}$  is the associated *pullback* metric. All the geometric quantities of the Riemannian analysis of  $\mathcal{M}$ , such as Christoffel’s symbols, covariant derivatives, curvature tensors, etc. can be defined. This is actually done in the following appendices of this paper.

Second, some important designations can be introduced. The metric  $\mathcal{G}_{ij}(\lambda)$  is a (*covariant*) *tensor* of rank-2, i.e. described by two subscript indices, on  $\mathcal{M}$ . This means that  $\mathcal{G}_{ij}$  satisfies a specific transformation under changes of coordinates in  $T_\lambda \Lambda$  such that the values of the bilinear form<sup>14</sup>  $\mathcal{G}_\lambda(\xi, \zeta) \triangleq \xi^i \zeta^j \mathcal{G}_{ij}(\lambda)$  that it induces are unmodified<sup>15</sup>. A function  $f : \Lambda \rightarrow \mathbb{R}$  is a *scalar field* on  $\mathcal{M}$ , or rank-0 tensor. A vector field  $\zeta^i(\lambda)$  on this manifold, which associates to each point  $\lambda$  a vector in the tangent plane  $T_\lambda \Lambda$ , is a function  $\zeta : \Lambda \rightarrow T_\lambda \Lambda \simeq \mathbb{R}^P$  also named (*contravariant*) rank-1 tensor, i.e. with one superscript. More generally, a rank- $(m, n)$  tensor is a quantity  $T_{j_1 \dots j_n}^{i_1 \dots i_m}(\lambda)$   $m$ -times contravariant and  $n$ -times covariant such that  $\mathcal{G}_{i_1 k_1} \dots \mathcal{G}_{i_m k_m} \xi_1^{k_1} \dots \xi_m^{k_m} T_{j_1 \dots j_n}^{i_1 \dots i_m}(\lambda) \zeta_1^{j_1} \dots \zeta_n^{j_n}$  is invariant under change of coordinates in  $T_\lambda \Lambda$  for any vectors  $\{\xi_1, \dots, \xi_m, \zeta_1, \dots, \zeta_n\}$  in this space.

## B Proof of Proposition 2

Let  $\gamma$  be a geodesic in  $\mathcal{M}$  with curvilinear parametrization, i.e. with  $|\gamma'(s)| = 1$ . Writing  $\gamma = \gamma(s)$  and  $\gamma' = \frac{d}{ds} \gamma(s)$ , we have  $\frac{d}{ds} g_\gamma(s) = \partial_i g_\gamma \gamma'^i$  and  $\frac{d^2}{ds^2} g_\gamma(s) = \partial_{ij} g_\gamma \gamma'^i \gamma'^j + \partial_k g_\gamma \gamma''^k$ , where we write abusively  $\partial_i g_\gamma = \partial_i g_\lambda|_{\lambda=\gamma(s)}$  and similarly for second order derivative.

We need now some elements of differential geometry. Since  $\gamma$  is a geodesic in  $\mathcal{M}$ , it respects the second order differential equation  $\gamma''^k + \Gamma_{ij}^k \gamma'^i \gamma'^j = 0$ , where the values  $\Gamma_{ij}^k = \frac{1}{2} \mathcal{G}^{lk} (\partial_j \mathcal{G}_{li} +$

<sup>14</sup>Also named first fundamental form [28].

<sup>15</sup>In the same way that the scalar product between two vectors in the usual Euclidean space is independent of the choice of coordinates.



$\partial_i \mathcal{G}_{jl} - \partial_l \mathcal{G}_{ij}$ ) are the Christoffel's symbols [28] derived from the metric  $\mathcal{G}_{ij}$ . Therefore, we get

$$rcl \frac{d^2}{ds^2} g_\gamma = \partial_{ij} g_\gamma \gamma'^i \gamma'^j - \partial_k g_\gamma \Gamma_{ij}^k \gamma'^i \gamma'^j \quad (15)$$

$$= \nabla_{ij} g_\gamma \gamma'^i \gamma'^j, \quad (16)$$

where  $\nabla_i g_\gamma = \partial_i g_\gamma$  and  $\nabla_{ij} g_\gamma = \nabla_i \nabla_j g_\gamma = \partial_{ij} g_\gamma - \partial_k g_\gamma \Gamma_{ij}^k$  are by definition the first order  $i$  and the second order  $ij$  covariant derivatives of  $g_\gamma$  respectively [28]. In addition, we can easily compute that for  $\mathcal{M} = (\Lambda, \mathcal{G}_{ij})$ ,

$$\Gamma_{ij}^k = \mathcal{G}^{kl} \langle \partial_{ij} g_\lambda, \partial_l g_\lambda \rangle. \quad (17)$$

The lower bound of the proposition comes simply from the projection of  $\frac{d^2}{ds^2} g_{\gamma(s)}$  onto  $g_\gamma$ . Indeed, for any  $\lambda \in \Lambda$ , since  $\|g_\lambda\|^2 = \langle g_\lambda, g_\lambda \rangle = 1$ ,  $\langle \partial_i g_\lambda, g_\lambda \rangle = 0$  and  $\langle \partial_{ij} g_\lambda, g_\lambda \rangle = -\mathcal{G}_{ij}$ . By (15),  $\langle \frac{d^2}{ds^2} g_{\gamma(s)}, g_\gamma \rangle = \langle \partial_{ij} g_\gamma, g_\gamma \rangle \gamma'^i \gamma'^j = -\mathcal{G}_{ij} \gamma'^i \gamma'^j = -1$ , and using Cauchy-Schwarz we get  $\|\frac{d^2}{ds^2} g_{\gamma(s)}\| \geq 1$ . Therefore, for  $\epsilon > 0$  and  $\gamma_\xi : [0, \epsilon] \rightarrow \Lambda$ , a segment of geodesic starting from  $\lambda$  with unit speed  $\xi$ ,

$$\mathcal{K} \geq \sup_{\xi: |\xi|=1} \left\| \frac{d^2}{ds^2} g_{\gamma_\xi(s)} \Big|_{s=0} \right\| \geq 1.$$

For the upper bound, coming back to any geodesic  $\gamma$ , we need to analyze directly  $\|\frac{d^2}{ds^2} g_{\gamma(s)}\|^2$ . Using (16) and the expression (17) of the Christoffel's symbols above, we have  $\|\frac{d^2}{ds^2} g_{\gamma(s)}\|^2 = \|\nabla_{ij} g_\gamma \gamma'^i \gamma'^j\|^2 \leq \langle \nabla_{ij} g_\gamma, \nabla_{kl} g_\gamma \rangle \mathcal{G}^{ik} \mathcal{G}^{jl}$ , where we used  $|\gamma'| = 1$  and the Cauchy-Schwarz (CS) inequality expressed in the Einstein's summation notation on rank-2 tensors. This latter states that, for the tensors  $A_{ij} = \nabla_{ij} g_\gamma$  and  $B^{ij} = \gamma'^i \gamma'^j$ ,  $|A_{ij} B^{ij}|^2 \leq |A_{ij} A_{kl} \mathcal{G}^{ki} \mathcal{G}^{lj}| |B^{ij} B^{kl} \mathcal{G}_{ki} \mathcal{G}_{lj}|$ , the equality holding if the two tensors are multiple of each other. We prove in [31] the general explanation for rank-n tensor as a simple consequence of the positive-definiteness of  $\mathcal{G}_{ij}$ .

Therefore, taking  $\gamma = \gamma_\xi$ , and since  $\gamma_\xi(0) = \lambda$ ,

$$\mathcal{K} \leq \sup_{\lambda \in \Lambda} \left[ \langle \nabla_{ij} g_\lambda, \nabla_{kl} g_\lambda \rangle \mathcal{G}^{ik} \mathcal{G}^{jl} \right]^{\frac{1}{2}}. \quad (18)$$

In the companion Technical Paper [31], we prove that this inequality is also equivalent to

$$\mathcal{K} \leq \sup_{\lambda \in \Lambda} \left[ R(\lambda) + \|\Delta g_\lambda\|^2 \right]^{\frac{1}{2}},$$

where  $R$  is the scalar curvature of  $\mathcal{M}$ , i.e. the quantity  $R = R_{ijkl} \mathcal{G}^{ik} \mathcal{G}^{jl}$  contracted from the curvature tensor  $R_{iklm} = \frac{1}{2}(\partial_{kl} \mathcal{G}_{im} + \partial_{im} \mathcal{G}_{kl} - \partial_{km} \mathcal{G}_{il} - \partial_{il} \mathcal{G}_{km}) + \mathcal{G}_{np}(\Gamma_{kl}^n \Gamma_{im}^p - \Gamma_{km}^n \Gamma_{il}^p)$ , and  $\Delta g_\lambda = \mathcal{G}_{ij} \nabla^i \nabla^j g_\lambda$  is the Laplace-Beltrami operator applied on  $g_\lambda$ . The curvature  $R$  requires only the knowledge of  $\mathcal{G}_{ij}(\lambda)$  (and its derivatives), implying just one step of scalar products computations, i.e. integrations in  $L^2(X)$ .

The reader who does not want to deal with differential geometry can however get rid of the covariant derivatives of Equation (18) by replacing them by usual derivatives. This provides however a weaker bound. Indeed, using the expression (17) of the Christoffel's symbols, some easy calculation provides  $0 \leq \langle \nabla_{ij} g_\lambda, \nabla_{kl} g_\lambda \rangle \mathcal{G}^{ik} \mathcal{G}^{jl} = \langle \partial_{ij} g_\lambda, \partial_{kl} g_\lambda \rangle \mathcal{G}^{ik} \mathcal{G}^{jl} - a_{ijk} a_{lmn} \mathcal{G}^{il} \mathcal{G}^{jm} \mathcal{G}^{kn}$ , with  $a_{ijk} = \langle \partial_{ij} g_\lambda, \partial_k g_\lambda \rangle$ .

Therefore,  $\langle \nabla_{ij} g_\lambda, \nabla_{kl} g_\lambda \rangle \leq \langle \partial_{ij} g_\lambda, \partial_{kl} g_\lambda \rangle \mathcal{G}^{ik} \mathcal{G}^{jl}$ , from the positive definiteness of  $\mathcal{G}^{ij}$  and  $\mathcal{G}_{ij}$ . Indeed, if we write  $W^{ijklmn} = \mathcal{G}^{il} \mathcal{G}^{jm} \mathcal{G}^{kn}$ , and if we gather indices  $ijk$  and  $lmn$  in the two multi-indices<sup>16</sup>  $I = (i, j, k)$  and  $L = (l, m, n)$ ,  $W^{IL}$  can be seen as a 2-D matrix in  $\mathbb{R}^{P^3 \times P^3}$ . It is then easy to check that the  $P^3$  eigenvectors of  $W^{IL}$  are given by the  $P^3$  combinations of the product of three of the  $P$  eigenvectors of  $\mathcal{G}^{ij}$ , i.e. the covariant vectors  $\zeta_i$  respecting the equation  $\mathcal{G}^{ij} \zeta_j = \mu \delta^{ij} \zeta_j$  for a certain  $\mu = \mu(\zeta) > 0$ . The matrix  $\mathcal{G}^{ij}$  being positive, the eigenvalues of  $W^{IL}$  are thus all positive, and  $W^{IL}$  is positive. Therefore,  $a_I W^{IL} a_L \geq 0$  for any tensor  $a_I = a_{ijk}$ . ■

## C Proof of Lemma 2

Recall that we use the gradient ascent defined from the optimization function  $\phi_r$  such that  $\phi_{r+1}(k) = \phi_r(k) + t_r \xi_r(k)$ , for a sequence of positive step size  $t_r$  increasing  $S_u$  at each step, and for a step direction  $\xi_r^i(\lambda) \triangleq |\nabla S_u(\phi_r(\lambda))|^{-1} \nabla^i S_u(\phi_r(\lambda))$ . From this definition, starting from  $k \in \Lambda$ , if  $\lim_{r \rightarrow +\infty} \phi_r(k) = k^\infty \in \Lambda$  exists, then  $k^\infty$  is a point where  $\nabla^i S_u(k^\infty) = 0$  for all  $i$ , since  $S_u(\phi_{r+1}(k)) = S_u(\phi_r(k)) + t_r |\partial S_u(\phi_r(k))| + O(t_r^2)$ .

How may the trajectory  $\mathcal{T}_k = \{\phi_r(k) : r \in \mathbb{N}\}$  contain a point  $\lambda'$  satisfying (14) ? Let us write  $\gamma_r(s)$  for the geodesic linking  $\phi_r(k)$  to  $\lambda_M$ , and define the *distance function*  $\zeta_r = d_{\mathcal{G}}(\lambda_M, \phi_r(k))$ . We have thus  $\gamma_r(0) = \phi_r(k)$  and  $\gamma_r(\zeta_r) = \lambda_M$ , where  $\lambda_M$  is the global maximum of  $S_u$ .

Case 1. If  $\xi_0^i \gamma_0'^j(0) \mathcal{G}_{ij}(k) < 0$ , i.e. the optimization starts in the wrong direction. The function  $\psi(s) = S_u(\gamma_0(s))$  is twice differentiable over  $[0, \zeta_0]$  and for  $s$  close to zero, we have  $\psi(0) > \psi(s)$  since  $\psi'(0) = \partial_i S_u(k) \gamma_0'^i(0) = |\nabla S_u(k)| \xi_0^i \gamma_0'^j(0) \mathcal{G}_{ij}(k) < 0$ .

Since  $\lambda_M$  is a global maximum of  $S_u$ ,  $\psi(0) < \psi(\zeta_0) = S_u(\lambda_M)$ . Therefore, there exists a  $s^* \in (0, \zeta_0)$  that minimizes  $\psi$ , i.e.  $\psi'(s^*) = 0$  with  $\psi(s^*) < \psi(0)$ . For  $\lambda_* = \gamma_0(s^*)$ , this implies that  $\lambda_*$  is critical since  $\psi'(s^*) = \partial_i S_u(\lambda_*) \gamma_0'^i(s^*) = 0$ . From Lemma 1,  $S_u(\lambda_M) - S_u(\lambda_*) \leq \frac{1}{2} \|u\|^2 d_{\mathcal{G}}(\lambda_M, \lambda_*)^2 (1 + \mathcal{K}) < \frac{1}{2} \|u\|^2 d_{\mathcal{G}}(\lambda_M, k)^2 (1 + \mathcal{K})$ , since  $d_{\mathcal{G}}(\lambda_M, \lambda_*) < d_{\mathcal{G}}(\lambda_0, k)$ . Finally, for any  $\lambda' \in \mathcal{T}_k$ ,  $S_u(\lambda_M) - S_u(\lambda') \leq S_u(\lambda_M) - S_u(k)$ , and  $S_u(\lambda_M) - S_u(\lambda') \leq S_u(\lambda_M) - S_u(\lambda_*) \leq \frac{1}{2} \|u\|^2 d_{\mathcal{G}}(\lambda_M, k)^2 (1 + \mathcal{K})$ , since  $S_u(k) \geq S(\lambda_*)$ .

Case 2. If  $\xi_0^i \gamma_0'^j(0) \mathcal{G}_{ij}(k) = 0$ . We have right away  $\gamma_0'^i(0) \partial_i S_u(k) = 0$ , and  $k$  is a critical point in the direction  $\lambda_M$ . Lemma 1 applied on  $k$  gives  $S_u(\lambda_M) - S_u(k) \leq \frac{1}{2} \|u\|^2 d_{\mathcal{G}}(\lambda_M, k)^2 (1 + \mathcal{K})$ . Since  $S_u(\lambda_M) - S_u(\lambda') \leq S_u(\lambda_M) - S_u(k)$  for any  $\lambda' \in \mathcal{T}_k$ , Equation (14) holds.

Case 3. If  $\xi_0^i \gamma_0'^j(0) \mathcal{G}_{ij}(k) > 0$ . Let us analyze the behavior of the distance function  $\zeta_r$ .

Let us introduce the function  $d_M(\lambda) = d_{\mathcal{G}}(\lambda_M, \lambda)$ . As for the Euclidean space, it is easy to prove<sup>17</sup> that  $\nabla^i d_M(\lambda) = -\gamma^i(0)$  if  $\gamma$  is the geodesic linking  $\lambda = \gamma(0)$  to  $\lambda_M$ . Therefore, since  $\zeta_{r+1} = d_M(\phi_{r+1}(k))$ , a Taylor expansion of  $d_M(\lambda)$  around  $\lambda = \phi_r(k)$  provides

$$\zeta_{r+1} = \zeta_r - t_r \xi_r^i(k) \gamma_r'^j(0) \mathcal{G}_{ij}(\phi_r(k)) + O(t_r^2). \quad (19)$$

For  $r = 0$ , if  $t_0$  is sufficiently small,  $\zeta_1 < \zeta_0$  and  $\zeta_r$  has either a local minima on a particular step  $r_m > 0$ , or it decreases monotonically and converges to a value  $\zeta_\infty = \lim_{r \rightarrow \infty} \zeta_r < \zeta_0$ .

<sup>16</sup>This can be seen as a relabelling of the  $P^3$  combinations of values for  $ijk$  into  $P^3$  different one-number indices  $I$ .

<sup>17</sup>The interested reader will find a proof of this basic differential geometry result in the companion Technical Report [31].

(i)  $\zeta$  has a local minima  $\zeta_{r_m} < \zeta_0$  on  $r_m > 0$  : Then,  $\zeta_{r_m+1} > \zeta_{r_m}$  and, using (19) with some implicit dependences,  $\zeta_{r_m+1} - \zeta_{r_m} = -t_{r_m} \gamma'_{r_m}(0) \xi_{r_m}^j \mathcal{G}_{ij} + O(t_{r_m}^2)$ . Therefore, for a sufficiently small step  $t_{r_m}$ ,  $\gamma'_{r_m}(0) \xi_{r_m}^j \mathcal{G}_{ij} < 0$  and we are in the same hypothesis as *Case 1* with the point  $\lambda' = \phi_{r_m}(k) \in \mathcal{T}_k$  instead of  $k$ . We obtain then  $S_u(\lambda_M) - S_u(\lambda') \leq \frac{1}{2}\|u\|^2 d_{\mathcal{G}}(\lambda_M, \lambda')^2 (1 + \mathcal{K}) < \frac{1}{2}\|u\|^2 d_{\mathcal{G}}(\lambda_M, k)^2 (1 + \mathcal{K})$ , since  $d_{\mathcal{G}}(\lambda_M, \lambda') = \zeta_{r_m} < \zeta_0 = d_{\mathcal{G}}(\lambda_M, k)$ .

(ii) If  $\zeta_r$  decreases monotonically for  $r > 0$  : Since  $\zeta_r \geq 0$ , the limit  $\lim_{r \rightarrow \infty} \zeta_r$  exists and converges to  $\zeta_\infty < \zeta_0$ . However, it is not guaranteed that the sequence  $\{\phi_r(k)\}$  converges to a point of  $\Lambda$ . Fortunately, since for all  $r > 0$ ,  $\phi_r(k)$  remains in the finite volume  $V_0 = \{\lambda \in \Lambda : d_M(\lambda) \leq d_M(k)\}$ , this sequence is bounded in the finite dimensional space  $\Lambda$ . Therefore, from the Bolzano-Weierstrass theorem on the metric space  $(\Lambda, d_{\mathcal{G}}(\cdot, \cdot))$ , we can find a convergent subsequence  $\{r_i \in \mathbb{N} : r_{i+1} > r_i\}$  such that  $\lim_{i \rightarrow \infty} \phi_{r_i}(k) = k_\infty \in V_0$ . On this point, we will have  $\nabla^i S_u(k_\infty) = 0$  for all  $i$ . So,  $k_\infty$  is an umbilical point and, from Lemma 1,

$$S_u(\lambda_M) - S_u(k_\infty) \leq \frac{1}{2}\|u\|^2 d_{\mathcal{G}}(\lambda_M, k_\infty)^2 (1 + \mathcal{K}).$$

From now on, we abuse notation and write  $\phi_{r_i}(k) = \phi_i(k)$ . Since  $\zeta_\infty^2 = d_{\mathcal{G}}(\lambda_M, k_\infty)^2 < d_{\mathcal{G}}(\lambda_M, k)^2 = \zeta_0^2$ , we can find a  $\delta > 0$  such that  $d_{\mathcal{G}}(\lambda_M, k_\infty)^2 + \delta < d_{\mathcal{G}}(\lambda_M, k)^2$ . Therefore, because  $\lim_{i \rightarrow \infty} S_u(\phi_i(k)) = S_u(k_\infty)$  by continuity of  $S_u$ , and since  $S_u(\phi_i(k))$  increases monotonically with  $i$ , there exists a  $i' > 0$  such that  $S_u(k_\infty) - S(\phi_{i'}(k)) \leq \frac{1}{2}\|u\|^2 \delta (1 + \mathcal{K})$ . With  $\lambda' = \phi_{i'}(k) \in \mathcal{T}_k$ , we finally get  $S_u(\lambda_M) - S_u(\lambda') < S_u(\lambda_M) - S_u(k_\infty) + \frac{1}{2}\|u\|^2 \delta (1 + \mathcal{K})$ , so that  $S_u(\lambda_M) - S_u(\lambda') \leq \frac{1}{2}\|u\|^2 (d_{\mathcal{G}}(\lambda_M, k_\infty)^2 + \delta) (1 + \mathcal{K}) < \frac{1}{2}\|u\|^2 d_{\mathcal{G}}(\lambda_M, k)^2 (1 + \mathcal{K})$ . This gives the result and concludes the proof.  $\blacksquare$

## References

- [1] S. Chen, D. Donoho, and M. Saunders, “Atomic decomposition by basis pursuit,” *SIAM Journ. Sci. Comp.*, vol. 20, no. 1, pp. 33–61, August 1998.
- [2] S. Mallat and Z. Zhang, “Matching pursuit with time-frequency dictionaries,” *IEEE T. Signal. Proces.*, vol. 41, no. 12, pp. 3397–3415, December 1993.
- [3] J. Tropp, “Greed is good: algorithmic results for sparse approximation,” *IEEE T. Inform. Theory.*, vol. 50, no. 10, pp. 2231–2242, October 2004.
- [4] E. Le Pennec and S. Mallat, “Sparse geometric image representations with bandelets,” *IEEE T. Image. Process.*, vol. 14, no. 4, pp. 423–438, April 2005.
- [5] D. Donoho and X. Huo, “Uncertainty principles and ideal atom decomposition,” *IEEE T. Inform. Theory.*, vol. 47, no. 7, pp. 2845–2862, November 2001.
- [6] R. Neff and A. Zakhori, “Very low bit rate video coding based on matching pursuits,” *IEEE T. Circ. Syst. Vid.*, vol. 7, no. 1, pp. 158–171, February 1997.
- [7] M. Goodwin and M. Vetterli, “Matching pursuit and atomic signal models based on recursive filterbanks,” *IEEE T. Signal. Proces.*, vol. 47, no. 7, pp. 1890–1902, July 1999.

- [8] P. Durka, D. Ircha, and K. Blinowska, "Stochastic time-frequency dictionaries for matching pursuit," *IEEE T. Signal. Proces.*, vol. 49, no. 3, pp. 507–510, March 2001.
- [9] R. Gribonval and E. Bacry, "Harmonic decomposition of audio signals with matching pursuit," *IEEE T. Signal. Proces.*, vol. 51, no. 1, pp. 101–111, January 2003.
- [10] R. Gribonval and P. Vandergheynst, "On the exponential convergence of matching pursuits in quasi-incoherent dictionaries," *IEEE T. Inform. Theory.*, vol. 52, no. 1, pp. 255–261, January 2006.
- [11] O. Divorra Escoda, L. Granai, and P. Vandergheynst, "On the use of a priori information for sparse signal approximations," *IEEE T. Signal. Proces.*, vol. 54, no. 9, pp. 3468–3482, September 2006.
- [12] A. Rahmoune, P. Vandergheynst, and P. Frossard, "Flexible motion-adaptive video coding with redundant expansions," *IEEE T. Circ. Syst. Vid.*, vol. 16, no. 2, pp. 178–190, February 2006.
- [13] C. De Vleeschouwer and A. Zakhor, "In-loop atom modulus quantization for matching pursuit and its application to video coding," *IEEE T. Image. Process.*, vol. 12, no. 10, pp. 1226–1242, October 2003.
- [14] R. Gribonval, "Fast matching pursuit with a multiscale dictionary of Gaussian chirps," *IEEE T. Signal. Proces.*, vol. 49, no. 5, pp. 994–1001, May 2001.
- [15] R. Figueras i Ventura, P. Vandergheynst, and P. Frossard, "Low-rate and flexible image coding with redundant representations," *IEEE T. Image. Process.*, vol. 15, no. 3, pp. 726–739, March 2006.
- [16] C. De Vleeschouwer and B. Macq, "Subband dictionaries for low-cost matching pursuits of video residues," *IEEE T. Circ. Syst. Vid.*, vol. 9, no. 7, pp. 984–993, October 1999.
- [17] P. Czerepinski, C. Davies, N. Canagarajah, and D. Bull, "Matching pursuits video coding: dictionaries and fast implementation," *IEEE T. Circ. Syst. Vid.*, vol. 10, no. 7, pp. 1103–1115, October 2000.
- [18] R. Neff and A. Zakhor, "Matching pursuit video coding. part i: Dictionary approximation," *IEEE T. Circ. Syst. Vid.*, vol. 12, no. 1, pp. 13–26, January 2002.
- [19] Y.-T. Chau, W.-L. Hwang, and C.-L. Huang, "Gain-shape optimized dictionary for matching pursuit video coding," *Signal Processing*, vol. 83, pp. 1937–1943, September 2003.
- [20] P. Schmid-Saugeon and A. Zakhor, "Dictionary design for matching pursuit and application to motion compensated video coding," *IEEE T. Circ. Syst. Vid.*, vol. 14, no. 6, pp. 880–886, June 2004.
- [21] M. Wakin, D. Donoho, H. Choi, and R. Baraniuk, "The multiscale structure of non-differentiable image manifolds," in *Wavelets XI. Proceedings of the SPIE*, vol. 5914, San Diego, CA, August 2005, pp. 413–429.

- [22] D. Donoho and C. Grimes, "Image Manifolds which are Isometric to Euclidean Space," *Journal of Mathematical Imaging and Vision*, vol. 23, no. 1, pp. 5–24, grimes 2005.
- [23] S. Amari, "Differential Geometry of Curved Exponential Families-Curvatures and Information Loss," *The Annals of Statistics*, vol. 10, no. 2, pp. 357–385, June 1982.
- [24] G. Watson and K. Gilholm, "Signal and image feature extraction from local maxima of generalised correlation," *Pattern Recognition*, vol. 31, no. 11, pp. 1733–1745, November 1998.
- [25] I. Todic, P. Frossard, and P. Vandergheynst, "Progressive coding of 3-d objects based on overcomplete decompositions," *IEEE T. Circ. Syst. Vid.*, vol. 16, no. 11, pp. 1338–1349, November 2006.
- [26] I. Bogdanova, P. Vandergheynst, and J.-P. Gazeau, "Continuous wavelet transform on the hyperboloid," *Appl. Comput. Harmon. Anal.*, 2005, (accepted).
- [27] S. Lang, *Differential manifolds*. Addison-Wesley Reading, Mass, 1972.
- [28] M. Carmo, *Riemannian Geometry*. Birkhauser, 1992.
- [29] R. Gribonval and P. Vandergheynst, "On the exponential convergence of matching pursuits in quasi-incoherent dictionaries," *IEEE T. Inform. Theory*, vol. 52, no. 1, pp. 255–261, January 2006.
- [30] S. Mallat, *A Wavelet Tour of Signal Processing*. Academic Press., 1998.
- [31] L. Jacques and C. De Vleeschouwer, "Discretization effects of continuous dictionary in matching pursuits: Density, convergence and optimization." UCL, Tech. Rep. TR-LJ-2007.01, July 2007, <http://www.tele.ucl.ac.be/~jacques/files/TR-LJ-2007.01.pdf> .
- [32] P. Niyogi, S. Smale, and S. Weinberger, "Finding the homology of submanifolds with high confidence from random samples," *Manuscript, Toyota Technological Institute, Chicago, Illinois*, 2004.
- [33] M. Davenport, M. Duarte, M. Wakin, J. Laska, D. Takhar, K. Kelly, and R. Baraniuk, "The smashed filter for compressive classification and target recognition," in *Computational Imaging V at SPIE Electronic Imaging*, San Jose, California, January 2007.
- [34] M. B. Wakin, "The geometry of low-dimensional signal manifolds," Ph.D. dissertation, Rice University, Houston, TX, USA, August 2006.
- [35] I. Daubechies, *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, 1992.
- [36] S. T. Ali, J.-P. Antoine, and J.-P. Gazeau, *Coherent States, Wavelets, and their Generalizations*. New York: Springer-Verlag, 2000.
- [37] O. Ferreira and P. Oliveira, "Subgradient Algorithm on Riemannian Manifolds," *Journal of Optimization Theory and Applications*, vol. 97, no. 1, pp. 93–104, April 1998.

- [38] D. Gabay, “Minimizing a differentiable function over a differential manifold,” *Journal of Optimization Theory and Applications*, vol. 37, no. 2, pp. 177–219, June 1982.
- [39] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.
- [40] R. Mahony and J. Manton, “The Geometry of the Newton Method on Non-Compact Lie Groups,” *Journal of Global Optimization*, vol. 23, no. 3, pp. 309–327, August 2002.
- [41] R. Figueras i Ventura, O. Divorra Escoda, and P. Vandergheynst, “A matching pursuit full search algorithm for image approximations,” EPFL, 1015 Ecublens, Tech. Rep., December 2004.
- [42] R. M. F. i Ventura, “Sparse image approximation with application to flexible image coding,” Ph.D. dissertation, Swiss Federal Institute of Technology, Lausanne, Switzerland, July 2005.
- [43] J.-P. Antoine, R. Murenzi, P. Vandergheynst, and S. Ali, *Two-dimensional Wavelets and Their Relatives*. Cambridge University Press, 2004.
- [44] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, “Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition,” in *In Proceedings of the 27th Annual Asilomar Conference on Signals, Systems and Computers*, November 1993.
- [45] E. Candes, J. Romberg, and T. Tao, “Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information,” *IEEE T. Inform. Theory.*, vol. 52, no. 2, pp. 489–509, 2006.
- [46] M. Duarte, M. Davenport, M. Wakin, and R. Baraniuk, “Sparse Signal Detection from Incoherent Projections,” in *IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP 2006 Proceedings.*, vol. 3, May 2006.
- [47] H. Rauhut, K. Schnass, and P. Vandergheynst, “Compressed sensing and redundant dictionaries,” *IEEE T. Inform. Theory.*, 2007, (submitted).
- [48] R. Baraniuk and M. Wakin, “Random projections of smooth manifolds,” *to appear in Foundations of Computational Mathematics*, 2007.