Average performance of OMP and Thresholding under dictionary mismatch

Marie-Christine Pali, Simon Ruetz, Karin Schnass
Department of Mathematics, University of Innsbruck
Email: firstname.lastname@uibk.ac.at

***Abstract*—This paper studies the performance of OMP in comparison to Thresholding in the case in which only a perturbed version of the generating dictionary is known. Both theory and numerical simulations show that simple Thresholding is a viable alternative to OMP in applications where only a perturbed version of the generating dictionary is given, such as dictionary learning.**

***Index Terms*—sparse approximation; Orthogonal Matching Pursuit; perturbed dictionary; average case analysis; dictionary learning; Thresholding**

I. Introduction

In dictionary learning the goal is to decompose a data matrix $\mathbf{Y} = (\mathbf{y}_1, \dots, \mathbf{y}_N) \in \mathbb{R}^{d\times N}$ into a dictionary matrix $\mathbf{\Phi} = (\boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_K) \in \mathbb{R}^{d\times K}$, where each column (also called atom) is normalised, i.e., $\|\boldsymbol{\phi}_k\|_2 = 1$, and a sparse coefficient matrix $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N) \in \mathbb{R}^{K\times N}$ such that $\mathbf{Y} \approx \mathbf{\Phi}\mathbf{X}$ and $\mathbf{X}$ is column-wise sparse. This problem can be formulated as an optimisation program

$$\min_{\mathbf{\Phi}\in\mathcal{D}_K} \sum_{n=1}^{N} \min_{\|\mathbf{x}_n\|_0\leq S} \|\mathbf{y}_n - \mathbf{\Phi}\mathbf{x}_n\|_2^2, \tag{1}$$

where $\mathcal{D}_K$ denotes the set of all dictionaries with K normalised atoms and $\|\mathbf{x}\|_0$ counts the number of non-zero entries of a vector $\mathbf{x}$. While sparse representations are important for performing many signal processing tasks such as denoising [1] or data reconstruction from incomplete information [2], solving a highly non-convex minimisation problem as in (1) is notoriously difficult [3]. Some of the most used dictionary learning algorithms belong to the class of alternating optimisation algorithms, which alternate between updating the sparse coefficient matrix $\mathbf{X}$ while fixing the dictionary $\mathbf{\Phi}$ and updating the dictionary $\mathbf{\Phi}$ while fixing $\mathbf{X}$. Popular examples include K-SVD [4], MOD [5], ITKrM [6] or the neural algorithms in [7]. Even updating the sparse coefficient matrix $\mathbf{X}$, meaning finding the best sparse approximation of each signal $\mathbf{y}_n$ in $\mathbf{\Phi}$, is generally NP-hard unless the dictionary forms an orthonormal system [8], [9]. In particular, in sparse approximation we want to approximate a given signal $\mathbf{y} \in \mathbb{R}^d$ by a linear combination of only a small number $S \ll d$ of elements $\boldsymbol{\phi}_i \in \mathbb{R}^d$ out of some given dictionary $\mathbf{\Phi} = (\boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_K)$. This means, denoting the restriction of $\mathbf{\Phi}$ and $\mathbf{x}$ to the columns resp. entries indexed by some set I by $\mathbf{\Phi}_I$ and $\mathbf{x}_I$, we want to find

$$\mathbf{y} \approx \sum_{k\in I} \phi_k x_k = \mathbf{\Phi}_I \mathbf{x}_I \quad \text{such that} \quad |I| = S \ll d. \tag{2}$$

The problem of finding the best S-sparse approximation of $\mathbf{y}$ in $\mathbf{\Phi}$, meaning the best S-support I and coefficient vector $\mathbf{x}$, however, is combinatorial. To approximate its solution efficiently, suboptimal routines that avoid searching through all possible sets I are typically used. One of the most practically used sparse approximation algorithms is Orthogonal Matching Pursuit (OMP) [10]. OMP finds the support iteratively by adding the index of the atom which has the largest absolute inner product with the residual and updating the residual. In particular, initialising with $\mathbf{r}_J = \mathbf{y}$ and $J = \emptyset$, it

$$\begin{aligned} \text{finds} \quad & i \in \arg\max_k |\langle \boldsymbol{\phi}_k, \mathbf{r}_J\rangle| \quad \text{and} \\ \text{updates} \quad & J \leftarrow J \cup \{i\} \quad \text{resp.} \quad \mathbf{r}_J = \mathbf{y} - P(\mathbf{\Phi}_J)\mathbf{y}, \end{aligned}$$

where $P(\mathbf{\Phi}_J)$ denotes the projection onto the span of atoms indexed by J, iterating until a stopping criterion is met. As the projection can be calculated iteratively the computational cost is determined by the K inner products $\langle \boldsymbol{\phi}_k, \mathbf{r}_J\rangle$ in each iteration, combining to an overall cost of $\mathcal{O}(SdK)$ [11].

On the other hand Thresholding or S-Thresholding with fixed sparsity level S finds the support by calculating

$$I \in \arg\max_{J:|J|=S} \|\mathbf{\Phi}_J^\top \mathbf{y}\|_1, \tag{3}$$

meaning it simply chooses those atoms which yield the S largest inner products in absolute value with the signal. Together with the projection this combines to a much reduced computational complexity of $\mathcal{O}(dK + S^3)$.

Over the years, quite a few results about sufficient conditions for OMP and Thresholding to recover the correct support emerged [12], [13]. Most recently, in [14] average case results for OMP were derived. However, these results assume exact knowledge of the generating dictionary, whereas in practice only an approximation might be available. Hence, they do not apply directly. In dictionary learning, for example, the initial guess (and also the subsequent updates) are naturally quite different from the signal generating dictionary. Thus results for Thresholding under dictionary mismatch can be found implicitly in several dictionary learning papers [6], [7], [15]. Further, a similar problem known as basis mismatch is studied in compressed sensing [16], [17]. There the dictionary rather

This work was supported by the Austrian Science Fund (FWF), Grant no. Y760.

than the signal coefficients is usually assumed to be random.
Contribution: In this work we provide a theoretical analysis of the average performance of OMP for the case in which we do not have the signal generating dictionary itself but only a perturbed version of it. Our theoretical results, confirmed by numerical simulations, indicate that Thresholding provides a viable and computationally cheaper alternative to OMP in case of dictionary mismatch. Finally, additional experiments show that on top of cost efficiency Thresholding also provides recovery advantages over OMP in dictionary learning.

II. Notation and setting

In the following we consider a dictionary $\mathbf{\Phi}$, i.e., a collection of K unit norm vectors $\boldsymbol{\phi}_i \in \mathbb{R}^d$, and define the coherence of $\mathbf{\Phi}$ as $\mu(\mathbf{\Phi}) := \max_{i\neq j} \left|\langle \boldsymbol{\phi}_i, \boldsymbol{\phi}_j\rangle\right|$. For a matrix $\mathbf{A} \in \mathbb{R}^{d\times K}$ we let $\mathbf{A}^\top$ denote the transpose of $\mathbf{A}$ and define $\|\mathbf{A}\|_{2,2} := \max_{\|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_2$. For a dictionary $\mathbf{\Phi}$ and an index set I of size S we define $\delta(\mathbf{\Phi}_I) = \|\mathbf{\Phi}_I^\top \mathbf{\Phi}_I - \mathbb{I}_S\|_{2,2}$. Note that if $\delta(\mathbf{\Phi}_I) < 1$ and therefore $\mathbf{\Phi}_I$ has full rank, we have for the projection $P(\mathbf{\Phi}_I) = \mathbf{\Phi}_I(\mathbf{\Phi}_I^\top \mathbf{\Phi}_I)^{-1}\mathbf{\Phi}_I^\top$.
For a perturbed version $\mathbf{\Psi}$ of a generating dictionary $\mathbf{\Phi}$, we set $\mathbf{Z} := \mathbf{\Phi} - \mathbf{\Psi}$ and define its distance to $\mathbf{\Phi}$ as $\varepsilon := \max_i \|\mathbf{z}_i\|_2$. The perturbation parameter $\nu := \max_{i,j} |\langle \boldsymbol{\psi}_i, \mathbf{z}_j\rangle|$ measures how correlated the perturbation of one atom is with the other perturbed atoms. Finally, for a vector $\mathbf{v} \in \mathbb{R}^K$ and an index ℓ, we define $\mathbf{v}_{\geq \ell} := \mathbf{v}_I$ for $I = \{\ell, \dots, K\}$.

Definition 1 (Signal model)**.** *Given a $d \times K$ dictionary $\mathbf{\Phi}$, we assume that our noisy signals are generated as*

$$\mathbf{y} = \mathbf{\Phi}_I \mathbf{x}_I + \boldsymbol{\eta} = \sum_{i\in I} \boldsymbol{\phi}_i \sigma_i c_{p(i)} + \boldsymbol{\eta}, \tag{4}$$

where $I \subset \{1, \cdots, K\}$ is a subset of size S chosen uniformly at random, p is some permutation satisfying $p(I) = \{1, \dots, S\}$ and $(\sigma_i)_i$ is a Rademacher sequence. The coefficients $\mathbf{c}$ are S-sparse and non-increasing, meaning $c_i = 0$ for $i > S$ and $c_i \geq c_{i+1}$ for $i \leq S$. The vector $\boldsymbol{\eta}$ denotes a sub-Gaussian noise vector with parameter ρ. In particular, this means that we have $\mathbb{E}(\boldsymbol{\eta}) = 0$ and for all vectors $\mathbf{v}$ with $\|\mathbf{v}\|_2 = 1$ and $\theta > 0$ the marginals $\langle \mathbf{v}, \boldsymbol{\eta}\rangle$ satisfy $\mathbb{E}(e^{\theta\langle \mathbf{v}, \boldsymbol{\eta}\rangle}) \leq e^{\theta^2\rho^2/2}$.

This signal model is quite general. Using Rademacher signs σ_i simply ensures that the coefficients x_i are centered, which together with boundedness is a quite common assumption [7], [18]. Further, we want to point out that sub-Gaussian noise includes both bounded and Gaussian noise. Choosing I uniformly at random among all sets of size S allows us to conclude that for any dictionary $\mathbf{\Psi}$ with small operator norm $\|\mathbf{\Psi}\|_{2,2}$ and small coherence $\mu(\mathbf{\Psi})$ we have $\delta(\mathbf{\Psi}_I) \leq 1/2$ with high probability [19], [20]. This could be replaced by a more general non-uniform sampling scheme, where similar conditions on $\mathbf{\Psi}$ including suitable weights again lead to $\delta(\mathbf{\Phi}_I) \leq 1/2$ with high probability [21].

III. Main results

Here we provide (partial) support recovery conditions for OMP and thresholding for the case in which the given input dictionary is not the signal generating dictionary but a perturbed version of it.

Theorem 1. *Assume the signals are generated following the model in (4) with signal generating dictionary $\mathbf{\Phi}$ and let $\mathbf{\Psi}$ be a perturbed version of $\mathbf{\Phi}$ with parameter ν.*
OMP: *Let $\ell \leq S$. If $\mathbf{\Psi}$ satisfies*

$$\mu(\mathbf{\Psi}) \leq \frac{1}{4n\log K} \quad \text{and} \quad \|\mathbf{\Psi}\|_{2,2}^2 \leq \frac{K}{16ne^2 S\log K}, \tag{5}$$

and for $\gamma \in (0,1)$ we have

$$\begin{aligned}\frac{1-\gamma}{2} > \mu(\mathbf{\Psi})\left(\max_{i\leq \ell}\frac{\|\mathbf{c}_{\geq i}\|_1}{c_i} + \sqrt{\ell}\max_{i\leq\ell}\cdot\frac{\|\mathbf{c}_{\geq i}\|_2}{c_i}\right) \\ + (1 + 2\ell\mu(\mathbf{\Psi}))\sqrt{2n\log K}\left(\frac{\nu\|\mathbf{c}\|_2 + \rho}{c_\ell}\right),\end{aligned} \tag{6}$$

then, except with probability $220K^{1-n}$, OMP using $\mathbf{\Psi}$ will recover a different atom from the support with coefficient size at least γc_ℓ in each of the first ℓ iterations.
Thresholding: *Let $\ell \leq S$. If for $\gamma \in (0,1)$ we have*

$$\frac{1-\gamma}{2} \geq \left(\mu(\mathbf{\Psi})\cdot\frac{\|\mathbf{c}\|_2}{c_\ell} + \frac{\nu\|\mathbf{c}\|_2 + \rho}{c_\ell}\right)\sqrt{2n\log K}, \tag{7}$$

then ℓ-Thresholding will recover ℓ atoms from the support with coefficient size at least γc_ℓ, except with probability $4K^{1-n}$.

Proof. We will show that OMP always picks a correct atom, whose coefficient is comparable to that with the largest coefficient still available. For J the current support we set $L := I \setminus J$ and let ℓ be the index of the largest remaining coefficient, i.e., $|x_\ell| = \|\mathbf{x}_L\|_\infty$. Further for $\gamma \in (0,1)$ we define $R := \{i \notin J : |x_i| < \gamma|x_\ell|\}$. We will show that for $\mathbf{r}_J = \mathbf{y} - P(\mathbf{\Psi}_J)\mathbf{y}$ we have

$$|\langle \boldsymbol{\psi}_\ell, \mathbf{r}_J\rangle| > \max_{i\in R} |\langle \boldsymbol{\psi}_i, \mathbf{r}_J\rangle|. \tag{8}$$

Rewriting $\mathbf{y} = \mathbf{\Phi}_I\mathbf{x}_I + \boldsymbol{\eta} = \mathbf{\Psi}_I\mathbf{x}_I + \mathbf{Z}_I\mathbf{x}_I + \boldsymbol{\eta}$, and abbreviating $Q(\mathbf{\Psi}_J) = \mathbb{I} - P(\mathbf{\Psi}_J)$ we get

$$\mathbf{r}_J = Q(\mathbf{\Psi}_J)\mathbf{\Psi}_L\mathbf{x}_L + Q(\mathbf{\Psi}_J)\mathbf{Z}_I\mathbf{x}_I + Q(\mathbf{\Psi}_J)\boldsymbol{\eta},$$

so in order to bound $|\langle \boldsymbol{\psi}_\ell, \mathbf{r}_J\rangle|$ from below, we need to bound the inner products of $\boldsymbol{\psi}_\ell$ with the terms on the r.h.s above. First note that by [19, Theorem 3.1] and (5) $\delta(\mathbf{\Psi}_J) \leq \delta(\mathbf{\Psi}_I) \leq 1/2$ except with probability $216K^{1-n}$. So for $\bar{L} = L\setminus\{\ell\}$ we have

$$\begin{aligned}&|\langle \boldsymbol{\psi}_\ell, \mathbf{\Psi}_L\mathbf{x}_L - P(\mathbf{\Psi}_J)\mathbf{\Psi}_L\mathbf{x}_L\rangle| \\ &\geq |x_\ell| - \left|\langle \mathbf{\Psi}_{\bar{L}}^\top\boldsymbol{\psi}_\ell, \mathbf{x}_{\bar{L}}\rangle\right| - \left|\langle \mathbf{\Psi}_J^\top\boldsymbol{\psi}_\ell, (\mathbf{\Psi}_J^\top\mathbf{\Psi}_J)^{-1}\mathbf{\Psi}_J^\top\mathbf{\Psi}_L\mathbf{x}_L\rangle\right| \\ &\geq \|\mathbf{x}_L\|_\infty - \|\mathbf{\Psi}_{\bar{L}}^\top\|_\infty\|\mathbf{x}_{\bar{L}}\|_1 \\ &\qquad - \|\mathbf{\Psi}_J^\top\boldsymbol{\psi}_\ell\|_2\|(\mathbf{\Psi}_J^\top\mathbf{\Psi}_J)^{-1}\|_{2,2}\|\mathbf{\Psi}_J^\top\mathbf{\Psi}_L\|_{2,2}\|\mathbf{x}_L\|_2 \\ &\geq \|\mathbf{x}_L\|_\infty - \mu(\mathbf{\Psi})\|\mathbf{x}_L\|_1 - \mu(\mathbf{\Psi})\sqrt{|J|}\frac{\delta(\mathbf{\Psi}_I)}{1-\delta(\mathbf{\Psi}_I)}\|\mathbf{x}_L\|_2.\end{aligned}$$

Analogue to above we get for $i \in R$

$$\begin{aligned}&|\langle \boldsymbol{\psi}_\ell, \mathbf{\Psi}_L\mathbf{x}_L - P(\mathbf{\Psi}_J)\mathbf{\Psi}_L\mathbf{x}_L\rangle| \\ &\qquad \leq \gamma\|\mathbf{x}_L\|_\infty + \mu(\mathbf{\Psi})\|\mathbf{x}_L\|_1 + \mu(\mathbf{\Psi})\sqrt{|J|}\|\mathbf{x}_L\|_2.\end{aligned}$$

Expanding again the projection we can bound the inner products of atoms with the perturbation term as

$$\begin{aligned}&\left|\langle \boldsymbol{\psi}_i, \mathbf{Z}_I\mathbf{x}_I - \boldsymbol{\Psi}_J(\boldsymbol{\Psi}_J^\top\boldsymbol{\Psi}_J)^{-1}\boldsymbol{\Psi}_J^\top\mathbf{Z}_I\mathbf{x}_I\rangle\right| \\ &\quad\le |\langle \boldsymbol{\psi}_i, \mathbf{Z}_I\mathbf{x}_I\rangle| + \frac{\|\boldsymbol{\Psi}_J^\top\boldsymbol{\psi}_i\|_2}{1-\delta(\boldsymbol{\Psi}_J)}\cdot\sqrt{|J|}\cdot\|\boldsymbol{\Psi}_J^\top\mathbf{Z}_I\mathbf{x}_I\|_\infty \\ &\quad\le \max_i |\langle \boldsymbol{\psi}_i, \mathbf{Z}_I\mathbf{x}_I\rangle|\cdot(1+2|J|\mu(\boldsymbol{\Psi}))\,. \end{aligned} \tag{9}$$

Since $x_j = c_{p(j)}\sigma_j$ we get via Hoeffding's inequality

$$\begin{aligned}&\mathbb{P}(|\langle \boldsymbol{\psi}_i, \mathbf{Z}_I\mathbf{x}_I\rangle| > t) = \mathbb{P}\big(|\textstyle\sum_j \langle \boldsymbol{\psi}_i, \mathbf{z}_j\rangle c_{p(j)}\sigma_j| > t\big) \\ &\quad\le 2\exp\left(\frac{-t^2}{2\sum_j\langle \boldsymbol{\psi}_i, \mathbf{z}_j\rangle^2 x_j^2}\right) \le 2\exp\left(\frac{-t^2}{2\nu^2\|\mathbf{x}_I\|_2^2}\right).\end{aligned}$$

Setting $t = t_\nu := \nu\|\mathbf{x}_I\|_2\sqrt{2n\log K}$ we get via a union bound that $\max_i |\langle \boldsymbol{\psi}_i, \mathbf{Z}_I\mathbf{x}_I\rangle| < t_\nu$ except with probability $2K^{1-n}$. Simply replacing $\mathbf{Z}_I\mathbf{x}_I$ by $\boldsymbol{\eta}$ in (9) we further get

$$|\langle \boldsymbol{\psi}_i, Q(\boldsymbol{\Psi}_J)\boldsymbol{\eta}\rangle| \le \max_i |\langle \boldsymbol{\psi}_i, \boldsymbol{\eta}\rangle|\cdot(1+2|J|\mu(\boldsymbol{\Psi}))\,.$$

Since $\boldsymbol{\eta}$ is sub-Gaussian, Markov's inequality leads to $\mathbb{P}(|\langle \boldsymbol{\psi}_i, \boldsymbol{\eta}\rangle| > t) \le 2e^{-t^2/(2\rho^2)}$. Setting $t = t_\rho := \rho\sqrt{2n\log K}$ and a union bound yield that $\max_i |\langle \boldsymbol{\psi}_i, \boldsymbol{\eta}\rangle| \le t_\rho$ except with probability $2K^{1-n}$.

After collecting all our bounds into (8) and rearranging, we get the following sufficient condition for OMP to pick another correct atom except with probability $220\cdot K^{1-n}$

$$\begin{aligned}\frac{1-\gamma}{2} &> \mu(\boldsymbol{\Psi})\left(\frac{\|\mathbf{x}_L\|_1}{\|\mathbf{x}_L\|_\infty} + \sqrt{|J|}\frac{\|\mathbf{x}_L\|_2}{\|\mathbf{x}_L\|_\infty}\right) \\ &\quad + \left(\frac{\nu\|\mathbf{x}_I\|_2+\rho}{\|\mathbf{x}_L\|_\infty}\right)(1+2|J|\mu(\boldsymbol{\Psi}))\sqrt{2n\log K}.\end{aligned}$$

To get the final result observe that $\|\mathbf{x}_I\|_2 = \|\mathbf{c}\|_2$ and that in the ℓ-th step $|J| = \ell - 1$ and $\|\mathbf{x}_L\|_\infty \ge c_\ell$. If $\|\mathbf{x}_L\|_\infty = c_i$ for the smallest possible i, then $\|\mathbf{x}_L\|_p \le \|\mathbf{c}_{\ge i}\|_p$ and

$$\frac{\|\mathbf{x}_L\|_p}{\|\mathbf{x}_L\|_\infty} \le \max_{i\le\ell}\frac{\|\mathbf{c}_{\ge i}\|_p}{c_i}.$$

To get the statement for thresholding, observe that

$$\langle \boldsymbol{\psi}_i, \mathbf{y}\rangle = \mathbf{x}_i + \langle \boldsymbol{\psi}_i, \boldsymbol{\Psi}_{I\setminus\{i\}}\mathbf{x}_{I\setminus\{i\}}\rangle + \langle \boldsymbol{\psi}_i, \mathbf{Z}_I\mathbf{x}_I\rangle + \langle \boldsymbol{\psi}_i, \boldsymbol{\eta}\rangle.$$

Hoeffding's inequality, the sub-Gaussianity of $\boldsymbol{\eta}$ and several union bounds, yield that except with probability $6K^{n-1}$

$$|\langle \boldsymbol{\psi}_i, \mathbf{y}\rangle| \le |\mathbf{x}_i| + (\mu(\boldsymbol{\Psi})\|\mathbf{x}\|_2 + \nu\|\mathbf{x}\|_2 + \rho)\sqrt{2n\log K},$$

for all i as well as the corresponding lower bound, so (7) ensures that the inner products of atoms having coefficients $c_i \ge c_\ell$ are larger than those having coefficients $c_i \le \gamma c_\ell$. □

IV. Comparison of OMP and Thresholding

In the perturbation and noise-free case our result reduces to that from [14], showing that the recovery condition for OMP becomes easier to fulfill if we have decaying coefficients. So for constant coefficients, we need $\mu(\boldsymbol{\Psi})S \lesssim 1$, while for coefficients forming a geometric sequence, meaning $c_i = \alpha^i$ for $\alpha \in (0,1)$ and $i \le S$, we only need $\mu(\boldsymbol{\Psi}) \lesssim 1-\alpha$ as well as $\mu^2(\boldsymbol{\Psi})S \lesssim 1-\alpha^2$ for full recovery. Unfortunately, in the case of perturbations, as ν grows, this advantage turns into a disadvantage, since the term $\|\mathbf{c}\|_2/c_S$ grows with faster decay, e.g. equaling $\sqrt{S}$ for constant coefficients and $\alpha^{-S}/\sqrt{1-\alpha^2}$ for the geometric sequence.

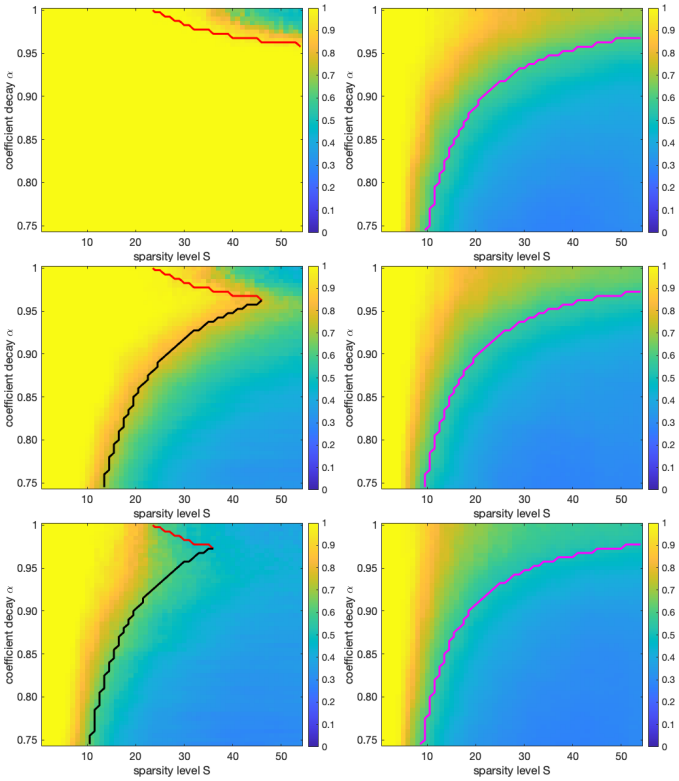

Fig. 1: Average percentage of correctly recovered atoms via OMP (left column) and Thresholding (right column) with perturbed dictionary $\boldsymbol{\Psi}$ where $\varepsilon = 0$ (top), $\varepsilon = 0.2$ (middle) and $\varepsilon = 0.5$ (bottom), for noiseless signals with generating dictionary $\boldsymbol{\Phi}$, various sparsity levels and coefficient decay parameters. The red, black and magenta lines indicate the theoretical decision boundaries.

For thresholding on the other hand the term scaled by ν, which grows with coefficient decay, already appears in the perturbation- and noise-free recovery-condition. This means that thresholding never performs well with large coefficient decay but also that its performance does not degrade dramatically with perturbations.

To better judge the influence of the perturbation parameter ν, we have a look at its extreme and typical size. For reasonable perturbation sizes, $\varepsilon := \max_k \|\mathbf{z}_k\|_2 \le 0.7$, we have $\nu = \max_{i,j}|\langle \boldsymbol{\psi}_i, \mathbf{z}_j\rangle| \approx \max_{i\ne j}|\langle \boldsymbol{\phi}_i, \mathbf{z}_j\rangle|$, so at worst, if $z_j \approx \varepsilon\boldsymbol{\phi}_k$, we have $\nu \approx \varepsilon$. On the other hand for random (rescaled Gaussian) perturbations we have $\nu \approx \varepsilon\sqrt{\log K/d}$. Also a more involved analysis – beyond the scope of this paper – for uniformly distributed supports, leads to a result corresponding to the above with $\nu \approx \|Z\|_{2,2}/\sqrt{K}$ [22].

To see how accurate our conditions are, we next conduct some numerical simulations in $\mathbb{R}^d$, for $d = 128$, for the case of

geometric coefficient sequences and random perturbations. We assume that the signals follow the model in (4), where the support I is chosen uniformly at random. For $\alpha \in [0.75, 1]$ and $S \in \{2, \dots, 54\}$ we set $c_i = \alpha^i$ for $i \leq S$ and $c_i = 0$ for all $i > S$. As generating dictionary $\mathbf{\Phi}$ we use the concatenation of the Dirac and DCT bases. We obtain a perturbed dictionary $\mathbf{\Psi}$ with distance ε to $\mathbf{\Phi}$ by setting $\boldsymbol{\psi}_k = \left(1 - \varepsilon^2/2\right) \boldsymbol{\phi}_k + \left(\varepsilon^2 - \varepsilon^4/4\right)^{1/2} \mathbf{v}_k$, where $\mathbf{v}_k$ is drawn uniformly at random from the unit sphere orthogonal to $\boldsymbol{\phi}_k$. For our experiments we use $N = 1000$ signals per sparsity level and decay parameter. The results in Figure 1 show that OMP outperforms Thresholding — but only for very small perturbations. This performance gap closes with growing levels of perturbation. In order to compare the sufficient conditions in Theorem 1 to our empirical results we plot the following boundaries

$$6 = \mu \cdot \left(\max_{i \leq S} \frac{\|\mathbf{c}_{\geq i}\|_1}{c_i} + \sqrt{S} \max_{i \leq S} \cdot \frac{\|\mathbf{c}_{\geq i}\|_2}{c_i} \right) \quad \text{(red)}$$

$$6 = \nu \cdot (1 + S\mu) \frac{\|\mathbf{c}\|_2}{c_S} \sqrt{\log K} \quad \text{(black)}$$

$$6 = (\nu + \mu) \cdot \frac{\|\mathbf{c}\|_2}{c_S} \sqrt{\log K} \quad \text{(magenta)}$$

for $\mu = \frac{1}{8} = \mu(\mathbf{\Phi}) \approx \mu(\mathbf{\Psi})$ and $\nu = \varepsilon/\sqrt{d}$. These results confirm the behaviour discussed above and show that the conditions in Theorem 1 are rather tight (up to constants).

V. Dictionary learning using OMP and Thresholding

Next we have a look at the implications of our results for the the motivating application of dictionary learning and compare the performance of Thresholding and OMP together with the atom update rules of K-SVD, MOD and ITKrM. We again generate signals in $\mathbb{R}^d$, for $d = 128$, using the concatenation of the Dirac and DCT bases as generating dictionary, meaning $K = 2d$ and $\mu(\mathbf{\Phi}) = 0.125$. We set $S = 6$, with the sparse coefficients forming a geometric sequence with decay factor $\alpha = 0.9$. This means $c_i = \kappa_S \alpha^i$ for $i \leq S$ and $c_i = 0$ for all $i > S$, where κ_S denotes some constant ensuring that $\|\mathbf{c}\|_2 = 1$. In case of noise, the noise vector is assumed to follow a normal distribution with variance $\rho_r^2 = (256d)^{-1}$, resulting in a signal to noise ratio of $\mathrm{SNR} = 256$. Each iteration uses $N = 20\,000$ fresh signals and the results are averaged over 10 runs.

As can be seen in Figure 2, all combinations of algorithms were able to fully recover the dictionary. Interestingly, the increased complexity of OMP does not seem to provide an advantage over Thresholding in the first few iterations. In the noiseless case OMP starts to outperform Thresholding only once the learned dictionary atoms are very close to their corresponding atoms in the generating dictionary, while in the noisy case, they perform nearly on par with each other. Taking into account that the Thresholding is computationally far less demanding than OMP there might not be a benefit in employing OMP in dictionary learning — in the early stages. Obviously an initialisation as defined in Section IV is quite unrealistic, which is why we repeat the same experiment using

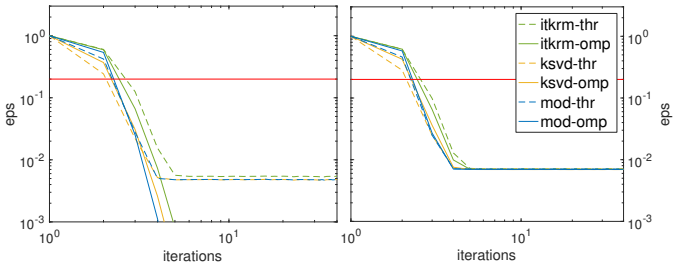

Fig. 2: Average distance of atoms to the generating dictionary for various dictionary learning algorithms using OMP (full lines) and Thresholding (dashed lines) for a well-behaved initialisation with $\varepsilon = 1$ (Section IV), using noiseless (left) resp. noisy training signals (right). The red line indicates the error at which the inner products between learned atoms and generating atoms would equal 0.98.

the noisy signals with a fully random initialisation. The results in Figure 3 paint a far more accurate picture of reality. It can be seen that, contrary to the previous experiment, OMP is not able to find all atoms of the generating dictionary (left), whereas Thresholding is able to find almost all atoms. Note that we used the convention that $\boldsymbol{\phi}_i$ is *found* if for a recovered $\boldsymbol{\psi}_k$ we have $|\langle \boldsymbol{\phi}_i, \boldsymbol{\psi}_k \rangle| \geq 0.99$, however the plot looks the same using 0.90 instead. Moreover, looking at the average distance of *found* atoms, we see that OMP is not able to outperform Thresholding either (right).

VI. Discussion

In this paper we have studied OMP and Thresholding in the case in which the generating dictionary is not known (or only a perturbed version of it is known). We compared sufficient conditions for OMP and Thresholding to find the correct support. It was shown that for small levels of perturbation, OMP does indeed outperform Thresholding, but that this gap closes with increasing levels of perturbation. This suggests that due to its computational efficiency Thresholding might be preferable to OMP in applications, where only an estimate of the generating dictionary is available, the prime example being dictionary learning.

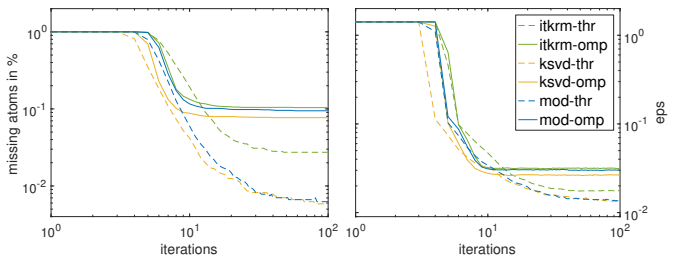

Fig. 3: (left) Percentage of atoms not found and (right) average distance to the generating atoms on *found* atoms, both using OMP (full lines) and Thresholding (dashed lines).

REFERENCES

[1] D. Donoho, M. Elad, and V. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Transactions on Information Theory*, vol. 52, no. 1, pp. 6–18, January 2006.

[2] J. Mairal, F. Bach, and J. Ponce, "Task-driven dictionary learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 791–804, 2012.

[3] A. M. Tillmann, "On the computational intractability of exact and approximate dictionary learning." *CoRR*, vol. abs/1405.6664, 2014.

[4] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing.*, vol. 54, no. 11, pp. 4311–4322, November 2006.

[5] K. Engan, S. Aase, and J. Husoy, "Method of optimal directions for frame design," in *ICASSP99*, vol. 5, 1999, pp. 2443 – 2446.

[6] K. Schnass, "Convergence radius and sample complexity of ITKM algorithms for dictionary learning." *Applied and Computational Harmonic Analysis*, vol. 45, no. 1, pp. 22–58, 2018.

[7] S. Arora, R. Ge, T. Ma, and A. Moitra, "Simple, efficient, and neural algorithms for sparse coding," in *COLT 2015 (arXiv:1503.00778)*, 2015.

[8] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness (Series of Books in the Mathematical Sciences)*. W. H. Freeman, 1979.

[9] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM J. Comput.*, vol. 24, no. 2, pp. 227–234, 1995.

[10] Y. Pati, R. Rezaiifar, and P. Krishnaprasad, "Orthogonal Matching Pursuit: recursive function approximation with application to wavelet decomposition," in *Asilomar Conf. on Signals Systems and Comput.*, 1993.

[11] R. Rubinstein, M. Zibulevsky, and M. Elad, "Efficient implementation of the K-SVD algorithm using batch orthogonal matching pursuit." CS Technion, Tech. Rep. 40(8), 2008.

[12] J. Tropp, "Greed is good: Algorithmic results for sparse approximation." *IEEE Transactions on Information Theory*, vol. 50, no. 10, pp. 2231–2242, October 2004.

[13] K. Schnass and P. Vandergheynst, "Average performance analysis for thresholding," *IEEE Signal Processing Letters*, vol. 14, no. 11, pp. 828–831, 2007.

[14] K. Schnass, "Average performance of Orthogonal Matching Pursuit (OMP) for sparse approximation," *IEEE Signal Processing Letters*, vol. 25, no. 12, pp. 1865–1869, 2018.

[15] M. Pali and K. Schnass, "Dictionary learning - from local towards global and adaptive," *arXiv:1804.07101*, 2021.

[16] Y. Chi, L. Scharf, A. Pezeshki, and R. Calderbank, "Sensitivity to basis mismatch in compressed sensing," *IEEE Transactions on Signal Processing*, vol. 59, no. 5, pp. 2182–2195, 2011.

[17] S. Bernhardt, R. Boyer, S. Marcos, and P. Larzabal, "Compressed sensing with basis mismatch: Performance bounds and sparse-based estimator," *IEEE Transactions on Signal Processing*, vol. 64, no. 13, pp. 3483–3494, 2016.

[18] N. Chatterji and P. L. Bartlett, "Alternating minimization for dictionary learning with random initialization," *Advances in Neural Information Processing Systems*, vol. 30, pp. 1997–2006, 2017.

[19] S. Chrétien and S. Darses, "Invertibility of random submatrices via tail-decoupling and matrix Chernoff inequality," *Statistics and Probability Letters*, vol. 82, pp. 1479–1487, 2012.

[20] J. Tropp, "Norms of random submatrices and sparse approximation," *Comptes Rendus Mathematique*, vol. 346, pp. 1271–1274, 2008.

[21] S. Ruetz and K. Schnass, "Submatrices with non-uniformly selected random supports and insights into sparse approximation," *SIAM Journal on Matrix Analysis and Applications*, vol. 42, no. 3, pp. 1268–1289, 2021.

[22] M.-C. Pali, "Dictionary learning & sparse modelling," Ph.D. dissertation, University of Innsbruck, 2021.