# Relaxed contractivity conditions for dictionary learning via Iterative Thresholding and K residual Means

Marie-Christine Pali and Karin Schnass
University of Innsbruck
{marie-christine.pali, karin.schnass}@uibk.ac.at

Alexander Steinicke
Montanuniversitaet Leoben
alexander.steinicke@unileoben.ac.at

## I. Introduction

Many tasks in high dimension signal processing, such as denoising or reconstruction from incomplete information, can be efficiently solved if the data at hand is known to be sparse in a dictionary $\Phi$, meaning given a $d\times K$ dictionary matrix, $\Phi = (\phi_1 \dots \phi_K) \in \mathbb{R}^{d\times K}$, with normalised columns also known as atoms, $\|\phi_k\|_2 = 1$, any data point $y$ can be approximately represented as superposition of $S$ atoms in the support $I$, that is $y \approx \sum_{i\in I} \phi_i x_i$ and $|I| = S$.

However, before being able to exploit this model for a given data class it is necessary to identify the parametrising dictionary, a process known as dictionary learning. By now, lots of algorithms exist that perform well in experiments and are popular in applications, [1], [2], [3] - for an introductory survey see for instance [4] - and also theoretical insights into dictionary learning are starting to accumulate. Unfortunately, so far the algorithms supported by global recovery results, [5], [6], [7], are rather unpractical, results for optimisation based approaches only hold for bases, [8], [9], [10], and results for practically usable alternating optimisation algorithms only guarantee local recovery, [11], [12], [13]. All theoretical results are based on rather stringent assumptions such as exact sparsity of the training data and knowledge of the sparsity level.

In this work we will improve on results in [13], [14] and give relaxed conditions that ensure that the alternating minimisation algorithm ITKrM contracts towards a generating dictionary.

## II. Dictionary learning via ITKrM

Given a batch of signals $Y = (y_1, \dots, y_N) \in \mathbb{R}^{d\times N}$, the goal of dictionary learning is to find a dictionary $\Phi$ yielding an approximately $S$-sparse representation of each signal $y_n$. That is, we want to decompose the data matrix into a dictionary $\Phi \in \mathbb{R}^{d\times K}$ and a sparse coefficient matrix $X = (x_1, \dots, x_N) \in \mathbb{R}^{K\times N}$, $Y \approx \Phi X$.

An algorithm designed to solve this problem is the Iterative Thresholding and K residual Means (ITKrM) algorithm, [13], summarised in Table I. As the name suggests, it alternates between updating the sparse coefficients based on the current dictionary and updating the dictionary based on the current coefficient matrix.

The algorithm has the advantage of working well on image data, compare Figures 1 and 2, while also being amenable to theoretical analysis. Given a generating dictionary $\Phi$, assume that the signals follow the model,

$$y = \textstyle\sum_i \phi_i c(p(i))\sigma_i \tag{1}$$

where $p$ is a permutation drawn uniformly at random, $\sigma$ is a Rademacher sequence, $r$ is sub-Gaussian noise and $c$ a non-increasing approximately $S$-sparse sequence, meaning $c(S+1)/c(S) \ll 1$.

Defining $d(\Phi,\Psi) := \max_k \|\phi_k - \psi_k\|_2^2$ (after reordering and sign flips), it was shown in [13] that ITKrM has a convergence radius of size $d(\Phi,\Psi) \leq \mathcal{O}(1/\sqrt{S})$ and in [14] that it contracts towards $\Phi$ for dictionaries as a far away as $d(\Phi,\Psi) \approx 2 - 2(\log K)^3/2\sqrt{S/d}$. The strategy used to prove both results relies in large part on bounding how often thresholding with an estimated dictionary $\Psi$ will fail to recover the generating support $I = \{p^{-1}(1), \dots, p^{-1}(S)\}$, and adding the resulting maximally possible error to each atom in the dictionary update. This is clearly suboptimal since the incorrect estimate of a generating support can at worst affect $2S$ atoms. Also the proof strategy cannot be extended to signal models with only approximately known sparsity level, where we do not have a large gap between the $S$ and $S{+}1$ largest coefficient, but only between the $S{-}T$ and $S{+}T{+}1$ largest coefficient for $T \geq 1$, $c(S{+}T{+}1)/c(S{-}T) \ll 1$. In such a situation we probably recover indices of the largest $S-T$ coefficients but not necessarily the full support. In other words, thresholding fails every time and adding the maximally possible error to each atom cannot lead to a non-zero convergence radius.

As a first step towards such a more general result, we therefore have to change the proof strategy and take into account that the failure of thresholding will only affect an atom $\phi_i$ if it is in the generating support, $i \in I$, but not in the thresholded support, $i \notin I^t$, or vice versa. Further, using more involved tools such as Freedman inequality, we get the following improvement over [14].

## III. Contraction of ITKrM

In the simplest case of noiseless, exactly $S$-sparse signals with balanced coefficients, meaning $c(i) \approx S^{-1/2}$ for $i \leq S$ and $c(i) = 0$ for $i > S$, our result reads as:

**Theorem III.1.** *Assume that our signals follow the model in* (1) *and that current dictionary estimate has coherence* $\mu(\Psi) := \max_{i\neq j} |\langle \psi_i, \psi_j \rangle|$, *and satisfies*

$$\mu(\Psi) \lesssim \frac{1}{\log K} \qquad \text{and} \qquad \|\Psi\|_{2,2}^2 \lesssim \frac{K}{S \log K}.$$

1) *If* $0 \leq d_{\mathit{min}} \leq d(\Psi,\Phi) \lesssim \frac{1}{\sqrt{\log K}}$
2) *or* $d(\Psi,\Phi) \gtrsim \frac{1}{\sqrt{\log K}}$ *but the cross Gram matrix* $\Phi^\star \Psi$ *is diagonally dominant in the sense that for all* $k$

$$|\langle \phi_k, \psi_k \rangle| \gtrsim \max\{\hat{\mu} \log K, \|\Psi\|_{2,2}\sqrt{S \log K/K)}\},$$

*where* $\hat{\mu} := \max_{i\neq j} |\langle \phi_i, \psi_j \rangle|$ *and* $d_{\mathit{min}}$ *depends on the number of training signals* $N \gtrsim S^2 K \log K$ *as well as the coherence* $\mu(\Phi)$ *and operator norm* $\|\Phi\|_{2,2}$ *of the generating dictionary, then with high probability one iteration of ITKrM will reduce the distance by at least a factor* $\kappa < 1$, *that is,* $d(\bar{\Psi},\Phi) < \kappa \cdot d(\Psi,\Phi)$.

## IV. Conclusion

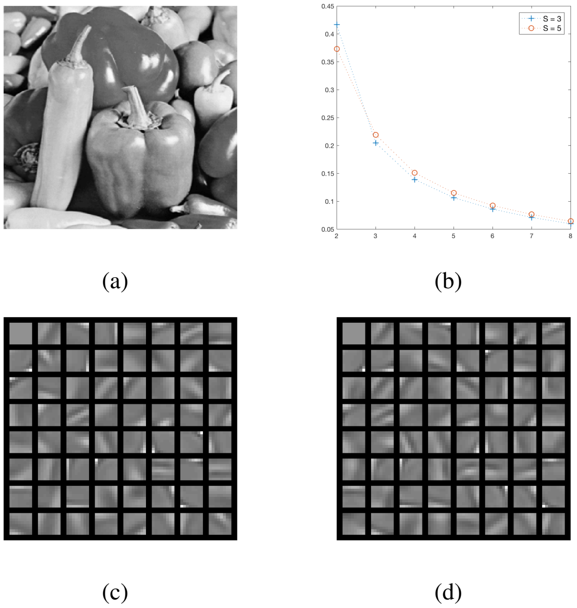
We have derived relaxed contraction conditions for ITKrM, that are satisfiable for dictionaries with distances to the generating dictionaries up to $d(\Phi,\Psi) \approx 2 - 2\sqrt{S \log K/d}$. The proof strategy further opens up the road for proving convergence results for more realistic signal models with less accurately known sparsity level, that better reflect practical situations, as for instance in Figures 1 and 2.

## Acknowledgments

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

```
ITKrM(Ψ, Y, S) - (one iteration) Given an input dictionary $\Psi$, a sparsity level $S$ and $N$ training signals $y_n$, do:
  • For all $n \in \{1, \dots, N\}$
      – For all $n$ find $I^t_{\Psi,n} = \operatorname{argmax}_{I:|I|=S} \|\Psi_I^\star y_n\|_1$
      – Set $a_n = y_n - P(\Psi_{I^t_{\Psi,n}})y_n$
  • For all $k \in I^t_{\Psi,n}$ calculate
      $\bar\psi_k = \frac{1}{N}\sum_n [a_n + P(\psi_k)y_n] \cdot \operatorname{sign}(\langle \psi_k, y_n\rangle) \cdot \chi(I^t_{\Psi,n}, k)$
  • Output $\bar\Psi = (\bar\psi_1/\|\bar\psi_1\|_2, \dots, \bar\psi_K/\|\bar\psi_K\|_2)$
```

TABLE I
ITKrM algorithm



(a) (b)

(c) (d)

Fig. 1. Two bases learned by ITKrM with sparsity $S = 3$ (c) and $S = 5$ (d) on all patches of *Peppers* (a). The training data consists of all patches after removing their mean (corresponding to projection on the orthogonal complement of the constant atom, left upper corner) with dimension $d = 63$. We then compute the average over all patches of the absolute values of the coefficients, that are calculated by OMP using all 64 atoms, sorted in descending order and divided by the patch norm, (b) (only some of the coefficients are plotted).

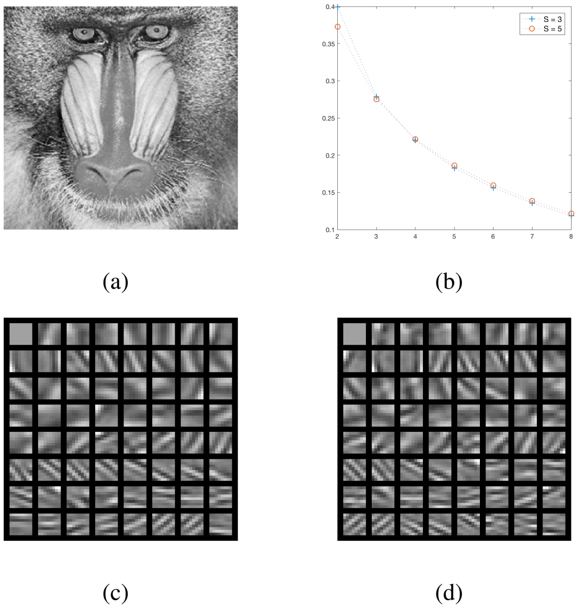
## References

[1] K. Kreutz-Delgado and B. Rao, "FOCUSS-based dictionary learning algorithms," in *SPIE 4119*, 2000.

[2] M. S. Lewicki and T. J. Sejnowski, "Learning overcomplete representations," *Neural Computations*, vol. 12, no. 2, pp. 337–365, 2000.

[3] 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.

[4] R. Rubinstein, A. Bruckstein, and M. Elad, "Dictionaries for sparse representation modeling," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1045–1057, 2010.



(a) (b)

(c) (d)

Fig. 2. Two bases learned by ITKrM with sparsity $S = 3$ (c) and $S = 5$ (d) on all patches of *Mandrill* (a). The training data consists of all patches after removing their mean with dimension $d = 63$. We then compute the average over all patches of the absolute values of the coefficients, that are calculated by OMP using all 64 atoms, sorted in descending order and divided by the patch norm, (b).

[5] A. Agarwal, A. Anandkumar, and P. Netrapalli, "Exact recovery of sparsely used overcomplete dictionaries," in *COLT 2014 (arXiv:1309.1952)*, 2014.

[6] S. Arora, R. Ge, and A. Moitra, "New algorithms for learning incoherent and overcomplete dictionaries," in *COLT 2014 (arXiv:1308.6273)*, 2014.

[7] B. Barak, J. Kelner, and D. Steurer, "Dictionary learning and tensor decomposition via the sum-of-squares method," in *STOC 2015 (arXiv:1407.1543)*, 2015.

[8] J. Sun, Q. Qu, and J. Wright, "Complete dictionary recovery over the sphere I: Overview and geometric picture," *IEEE Transactions on Information Theory*, vol. 63, no. 2, pp. 853–884, 2017.

[9] ——, "Complete dictionary recovery over the sphere II: Recovery by Riemannian trust-region method," *IEEE Transactions on Information Theory*, vol. 63, no. 2, pp. 885–915, 2017.

[10] N. Chatterji and P. Bartlett, "Alternating minimization for dictionary learning with random initialization," *arXiv:1711.03634*, 2017.

[11] K. Schnass, "On the identifiability of overcomplete dictionaries via the minimisation principle underlying K-SVD," *Applied and Computational Harmonic Analysis*, vol. 37, no. 3, pp. 464–491, 2014.

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

[13] 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.

[14] ——, "Dictionary learning - from local towards global and adaptive," *arXiv:1804.07101*, 2018.