Dictionary Learning
& Sparse Modelling

Dissertation in Mathematics

submitted by

Marie-Christine Pali

to the Faculty of Mathematics, Computer Science
and Physics of the University of Innsbruck

in partial fulfilment of the requirements
for the degree of Doctor of Philosophy

Advisor: Univ.-Prof. Karin Schnass
Co-Advisors: Univ.-Prof. Matthias Meiners
Univ.-Prof. Justus Piater

April 20, 2021

Acknowledgements

First of all, I want to thank my supervisor Prof. Karin Schnass for the many things she taught me during my PhD, for the many opportunities she offered me and especially for her support (in all situations). I also want to thank my other supervisors for always having an open door for me.

A big thank you also goes to all my colleagues from the Department of Mathematics at the University of Innsbruck for the great working atmosphere. At this point, a very special thank you goes to Christian Bargetz and Friederike Waldner for many funny coffee breaks and for always being there for me.

Apart from the University of Innsbruck, I want to thank Andreas Kofler (in Berlin) for many fruitful discussions and a great cooperation. Further, a big thanks goes also to all the people I had the chance to work with during my PhD and especially to those who became friends.

Thanks to the Austrian Science Fund for the financial support via the FWF-START project Y760.

Finally, I want to thank the people without whom none of this would have been possible - my whole family. Thank you for all the support and providing me with continuous love. The last and one of the biggest thanks goes to one of the greatest people in my life - my grandpa, who was always there for me.

Abstract

One of the key findings on which many signal processing tasks are built is that even high-dimensional signals admit some kind of sparse representation in a suitable generator system. This means, given such a system, called a dictionary, a given signal can be approximated by a linear combination of only a few of these dictionary elements. In this thesis we have a closer look at algorithms for learning dictionaries from data, algorithms for the sparse approximation of signals as well as their use in real-world applications.

In the first part of this thesis we study the contractive behaviour of dictionary learning via the Iterative Thresholding and K residual Means (ITKrM) algorithm. In particular, we show that ITKrM is a contraction under much more relaxed conditions than previously thought necessary and further analyse situations in which ITKrM does not recover the signal generating dictionary. Based on the insights gained, we develop a replacement strategy that allows ITKrM to escape from spurious fixed points towards the generating dictionary. Further, we introduce a strategy to learn dictionaries without the knowledge of the sparsity level and the dictionary size, leading to an adaptive version of ITKrM.

In the second part of this thesis we present an application where we use dictionary learning and sparse approximation algorithms for the reconstruction of highly undersampled MR images. In several experiments we show the competitiveness and advantages of the adaptive version of ITKrM compared to other well-established methods. We also conduct experiments to show the importance of the adaptive choice of the sparsity level and the dictionary size.

The last part of this thesis is devoted to the question how sparse approximation algorithms perform in situations where the given dictionary is not the same as the signal generating dictionary but a perturbed version of it. This occurs for example when they are used within dictionary learning algorithms. For that, we provide average case results for one specific sparse approximation algorithm - Orthogonal Matching Pursuit (OMP) - in the presence of perturbations of the generating dictionary and compare them with results obtained for thresholding, a computationally much lighter and simpler sparse approximation algorithm.

Contents

Appendix

List of Figures

List of Tables

Chapter 1

Introduction

The title of this thesis is Dictionary Learning & Sparse Modelling and therefore we want to start with a short explanation what dictionaries and sparsity are, why they are useful and the difficulties which may arise. In dictionary learning and sparse modelling we always work with signals, this means, vectors in some d-dimensional vector space. The concept which lays the foundation for dictionary learning and sparse modelling is sparsity. A vector or a matrix is called sparse if most of its entries are zero, and approximately sparse if most of its entries are very small. Sparse vectors (or matrices) are very important as they are easy to store and to compute with. For example, if we want to store a vector $y \in \mathbb{R}^d$ or a matrix $M \in \mathbb{R}^{m \times n}$, we normally have to remember d resp. mn numbers. However, if they are sparse with $S \ll d, (mn)$ non-zero entries, we only have to remember the non-zero components and their location, meaning $2S$ numbers. Similarly if we want to calculate with sparse vectors or matrices, we only have to perform operations between the non-zero components of the one with the other. This significantly reduces the number of calculations.

From this we see that sparsity is a great property. However, it seems to be too restrictive to be useful. Vectors or matrices are in general not sparse or can be sparsely represented (or approximated) in some orthonormal basis, meaning, can be written as a linear combination of only few of these basis elements. Even if we remove the requirements for orthogonality and consider any kind of basis, the concept of sparsity still seems to be too restrictive as there are not many classes of signals which are (approximately) sparse in such a basis. An example are images, which can be sparsely approximated in a wavelet basis. This means, if we want to store a lot of images y_n, knowing their sparse representation in such a basis Φ, we only have to store the non-zero coefficients and Φ. If the images are needed, they can be quickly reconstructed by multiplying the sparse coefficient vectors x_n with Φ, meaning $y_n = \Phi x_n$. This is for example used in the jpeg data compression standard.

However, for many classes of signals we in general do not have a basis in which they are sparse or approximately sparse. For that, we need something more general which is then called a dictionary. In case of having a class of d-dimensional signals which we want to sparsely approximate, a dictionary corresponds to a $d \times K$ matrix Φ with normalised columns, also referred to as atoms, and $K \geq d$ (or at least $K \geq \text{rank}(\Phi)$). Following the rule - the sparser the better - we often have $K \gg d$ since very sparse representations become more likely if we have more elements to build a signal. Such dictionaries are in general called overcomplete dictionaries, but since in this thesis we are mainly interested in the overcomplete case with $K > d$, we refer to them simply as dictionaries. The advantage of being able to better sparsely approximate signals however, comes with the drawback that these representations are not unique. This means, for each signal there exists more than one sparse representation in the dictionary and among these we want to find the sparsest one. Solving such problems is very difficult, even in the case where the dictionary forms a basis. For that, a lot of sparse approximation algorithms have been introduced as for example (O)MP ((Orthogonal) Matching Pursuit), BP (Basis Pursuit) or HTP (Hard Thresholding Pursuit), to name just a few. An assumption that all these algorithms have in common is that the sparsity providing dictionary is given. Such dictionaries can either be obtained by a careful analysis of the given data class, or even more efficiently, they can also be learned from data. The latter approach is called dictionary learning. Also for this problem there exist a lot of algorithms to choose from, as for example K-SVD (K Singular Value Decomposition), MOD (Method of Orthogonal Directions) or ITKrM (Iterative Thresholding and K residual Means).

Much research has been done in the direction of algorithmic development and also theoretical results have started to accumulate. Theoretical identification results are very important as they are needed to quantify the conditions on the dictionary, the coefficient model that generates the sparse signals and also the number of training signals needed for these algorithms to be successful. In this thesis, in Chapter 3, we will extend some existing recovery results for ITKrM and show that it behaves well under much more relaxed conditions than previously thought. Based on the insights gained there, in Chapter 4 we will introduce a strategy to further improve the convergence behaviour of ITKrM and to automatically choose the sparsity level and the dictionary size, which are needed to be given as input parameter to the algorithm.

Beside developing algorithms and analysing their theoretical performance, a lot of research has been done in the direction of how sparsity can be exploited for efficient data processing. For example, it has been shown that sparse signals are very robust to noise or corruption. Thus, by modelling signals as sparse linear combinations of some dictionary elements they can be easily denoised or restored from incomplete information. In Chapter 5 we will show an application where we use dictionary learning and sparse approximation algorithms for the reconstruction of MR (Magnetic Resonance) images from highly undersampled data.

Sparse approximation algorithms have not only been shown to work very well in practice but there exists also a substantial amount of theory concerning the analysis of their worst case or average case performance. The only drawback of most of these results is that they are based on the assumption that the signal generating dictionary is given. However, this may not hold true in all applications. For instance, sparse approximation algorithms are used within dictionary learning algorithms where (especially in the first iterations) the learned dictionary can be completely different from the signal generating dictionary. In particular, in Chapter 5 we will see that K-SVD and ITKrM produce comparable results. However, this seems quite surprising as K-SVD uses the more sophisticated sparse approximation algorithm OMP whereas ITKrM uses only simple thresholding. Therefore, in Chapter 6, we will analyse the average performance of OMP in case where we do not have the signal generating dictionary but only a perturbed version of it. We will also compare the results obtained for OMP with those obtained for thresholding.

1.1 Outline

This thesis is structured as follows. In Chapter 2 we describe the concepts of dictionary learning, sparse representations and approximations in more detail. We also introduce the main algorithms which are used in this thesis, meaning, ITKrM, K-SVD, OMP as well as the thresholding algorithm and discuss some related problems.

Chapter 3 studies the contractive behaviour of the ITKrM algorithm. After introducing and discussing some existing results, we provide a refined contraction theorem. In particular, we show that one iteration of ITKrM is a contraction under much more relaxed conditions than previously thought necessary.

In Chapter 4 we analyse situations in which ITKrM does not recover the generating dictionary. This will show us that there seem to exist stable fixed points which are not equivalent to the generating dictionary and have some very special structure. Based on an analysis of the residuals at these spurious fixed points, we develop a replacement strategy that allows ITKrM to escape towards the generating dictionary. With the help of the candidate atoms used for replacement we further introduce a strategy for the automatic choice of the sparsity level S and the dictionary size K, meaning, where S and K are adaptively chosen dependent on the data under consideration.

In Chapter 5 we present an application of dictionary learning and sparse approximation algorithms for the reconstruction of highly undersampled MR images. For dictionary learning we use the adaptive version of the ITKrM algorithm and propose an adaptive version of OMP, which we use for sparse coding. In various experiments we show their competitiveness and advantages against K-SVD+OMP as well as ITKrM+OMP. Compared to these methods, we show that the adaptive algorithms are significantly faster and at the same time highly facilitate the application in the clinical routine.

In Chapter 6 we analyse the average case performance of OMP in presence of perturbations of the generating dictionary. In particular, we provide conditions ensuring (partial-) support recovery for noiseless as well as noisy signals in case where the given input dictionary is not the signal generating dictionary itself but a perturbed version of it. We also compare the results obtained for OMP with those obtained for simple thresholding. The theoretical bounds are then illustrated by various numerical experiments.

In Chapter 7 we summarise and discuss the results of this thesis and point out further directions of research.

1.2 Notations

Before we start, we have to introduce some notations and definitions. In the following, usually subscripted letters will denote vectors with the exception of ε, α, ω, γ, λ, where they are numbers. For instance $x_n \in \mathbb{R}^K$ vs. $\varepsilon_k \in \mathbb{R}$, however, it should always be clear from the context what we are dealing with. For a vector $v \in \mathbb{R}^d$ and $1 \leq p < \infty$, we denote its p-norm by

$$\|v\|_p = \big(\sum_{i=1}^{d} |v_i|^p\big)^{\frac{1}{p}}$$

and for its maximum norm we write $\|v\|_\infty = \max_{i=1,\dots,d} |v_i|$. Similarly, for a matrix $M = (m_{ij})_{i,j} \in \mathbb{R}^{d \times n}$ we define the maximum p-norm of a column of M as

$$\|M\|_{1,p} = \max_{j=1,\dots,n} \big(\sum_{i=1}^{d} |m_{ij}|^p\big)^{\frac{1}{p}}$$

as well as for $p = \infty$, $\|M\|_{1,\infty} = \max_{i,j} |m_{ij}|$. We denote its operator norm by $\|M\|_{2,2} = \max_{\|x\|_2=1} \|Mx\|_2$ and its Frobenius norm by $\|M\|_F = \operatorname{tr}(M^\star M)^{1/2}$. Remember that we have $\|M\|_{2,2} \leq \|M\|_F$. For the matrix M we denote its (conjugate) transpose by $M^\star$ and its Moore-Penrose pseudo-inverse by $M^\dagger$.

We consider a **dictionary** Φ, a collection of K unit norm vectors $\phi_k \in \mathbb{R}^d$, $\|\phi_k\|_2 = 1$. By abuse of notation we will also refer to the $d \times K$ matrix collecting the atoms as its columns as the dictionary, that is, $\Phi = (\phi_1, \dots \phi_K)$. The maximal absolute inner product between two different atoms is called the **coherence** $\mu(\Phi)$ of a dictionary, $\mu(\Phi) = \max_{k \neq j} |\langle \phi_k, \phi_j \rangle|$.

By Φ_I we denote the restriction of the dictionary to the atoms indexed by I, that is, $\Phi_I = (\phi_{i_1}, \dots, \phi_{i_S})$, $i_j \in I$, and by $P(\Phi_I)$ the orthogonal projection onto the span of the atoms indexed by I, that is, $P(\Phi_I) = \Phi_I \Phi_I^\dagger$. Note that in case the atoms indexed by I are linearly independent we have $\Phi_I^\dagger = (\Phi_I^\star \Phi_I)^{-1} \Phi_I^\star$. We also define $Q(\Phi_I)$ to be the orthogonal projection onto the orthogonal complement of the span of Φ_I, that is, $Q(\Phi_I) = \mathbb{I}_d - P(\Phi_I)$, where $\mathbb{I}_d$ is the identity operator (matrix) in $\mathbb{R}^d$.

(Ab)using the language of compressed sensing we define $\delta_I(\Phi)$ as the smallest number such that all eigenvalues of $\Phi_I^\star \Phi_I$ are included in $[1-\delta_I(\Phi), 1+\delta_I(\Phi)]$ and the **isometry constant** $\delta_S(\Phi)$ of the dictionary as $\delta_S(\Phi) := \max_{|I|\leq S} \delta_I(\Phi)$. When clear from the context we will usually omit the reference to the dictionary. For more details on isometry constants see for instance [11].

For a (sparse) signal $y = \sum_k \phi_k x_k$ we will refer to the indices of the S coefficients with largest absolute magnitude as the S-support of y. Again, we will omit the reference to the sparsity level S if clear from the context.

To keep the sub(sub)scripts under control we denote the **indicator function of a set** $\mathcal{V}$ by $\chi(\mathcal{V}, \cdot)$, that is $\chi(\mathcal{V}, v)$ is one if $v \in \mathcal{V}$ and zero else. The set of the first S integers we abbreviate by $\mathbb{S} = \{1, \ldots, S\}$.

We define the **distance** of a dictionary Ψ to a dictionary Φ as

$$d(\Phi, \Psi) := \max_k \min_\ell \|\phi_k \pm \psi_\ell\|_2 = \max_k \min_\ell \sqrt{2 - 2|\langle \phi_k, \psi_\ell \rangle|}. \tag{1.1}$$

Note that this distance is not a metric since it is not symmetric. For example, if Φ is the canonical basis and Ψ is defined by $\psi_i = \phi_i$ for $i \geq 3$, $\psi_1 = (e_1 + e_2)/\sqrt{2}$, and $\psi_2 = \sum_i \phi_1/\sqrt{d}$ then we have $d(\Phi, \Psi) = 1/\sqrt{2}$ while $d(\Psi, \Phi) = \sqrt{2 - 2/\sqrt{d}}$. The advantage is that this distance is well defined also for dictionaries of different sizes. A **symmetric distance** between two dictionaries Φ, Ψ of the same size could be defined as the maximal distance between two corresponding atoms, that is,

$$d_s(\Phi, \Psi) := \min_{p \in \mathcal{P}} \max_k \|\phi_k \pm \psi_{p(k)}\|_2, \tag{1.2}$$

where $\mathcal{P}$ is the set of permutations of $\{1, \ldots, K\}$. The distances are equivalent whenever there exists a permutation p such that after rearrangement, the cross-Gram matrix $\Phi^\star \Psi$ is diagonally dominant, that is, $\min_k |\langle \phi_k, \psi_k \rangle| > \max_{k \neq j} |\langle \phi_k, \psi_j \rangle|$. Since the main assumption for our results will be such a diagonal dominance we will state them in terms of the easier to calculate asymmetric distance and assume that Ψ is already signed and rearranged in a way that $d(\Phi, \Psi) = \max_k \|\phi_k - \psi_k\|_2$. We then use the abbreviations $\alpha_{\min} = \min_k |\langle \phi_k, \psi_k \rangle|$ and $\alpha_{\max} = \max_k |\langle \phi_k, \psi_k \rangle|$. The maximal absolute inner product between two non-corresponding atoms will be called the **cross-coherence** $\mu(\Phi, \Psi)$ of the two dictionaries, $\mu(\Phi, \Psi) = \max_{k \neq j} |\langle \phi_k, \psi_j \rangle|$.

We will also use the following decomposition of a dictionary Ψ into a given dictionary Φ and a perturbation dictionary Z. If $d(\Psi, \Phi) = \varepsilon$ we set $\|\psi_k - \phi_k\|_2 = \varepsilon_k$, where by definition $\max_k \varepsilon_k = \varepsilon$. We can then find unit vectors z_k with $\langle \phi_k, z_k \rangle = 0$ such that

$$\psi_k = \alpha_k \phi_k + \omega_k z_k, \quad \text{for} \quad \alpha_k := 1 - \varepsilon_k^2/2 \quad \text{and} \quad \omega_k := (\varepsilon_k^2 - \varepsilon_k^4/4)^{\frac{1}{2}}. \tag{1.3}$$

Note that if the cross-Gram matrix $\Phi^\star \Psi$ is diagonally dominant we have $\alpha_{\min} = \min_k \alpha_k$, $\alpha_{\max} = \max_k \alpha_k$ and $d(\Psi, \Phi) = \sqrt{2 - 2\alpha_{\min}}$.

Chapter 2

Dictionaries and Sparsity

In this chapter we describe the concepts of dictionary learning, sparse signal representations and approximations in more detail. We also give a brief introduction to the main algorithms used in this thesis.

As already mentioned in the introduction, one way to handle high-dimensional data is by using the concept of sparsity. In particular, having a sparse representation of the signals of interest can be immensely practical as this significantly reduces the dimensionality of the signals. This not only reduces the number of values to be stored and calculations to be performed but also allows us to extract important features. A lot of signal processing tasks such as denoising [19], or data reconstruction from incomplete information [42, 47], can be efficiently performed when having a sparse representation or approximation of the signals of interest.
In concrete terms this means, given a dictionary $\Phi = (\phi_1, \dots, \phi_K) \in \mathbb{R}^{d \times K}$ with normalised columns, $\|\phi_k\|_2 = 1$, a signal $y \in \mathbb{R}^d$ has a S-sparse representation in Φ, or is called S-sparse in Φ, if there exists some index set I with $|I| = S \ll d$, such that we can write

$$y = \sum_{i \in I} \phi_i x_i = \Phi_I x_I. \tag{2.1}$$

The signal y has a S-sparse approximation in Φ, or is called approximately S-sparse in Φ, if there exists I with $|I| = S \ll d$, and ε small, $\|\varepsilon\|_2 \ll \|y\|_2$, such that we can write

$$y = \sum_{i \in I} \phi_i x_i + \varepsilon \approx \Phi_I x_I. \tag{2.2}$$

In classical sparsity research there are now two types of problems. The first one is concerned with how to find sparse approximations/representations given a sparsity inducing dictionary and the second one with how to exploit sparsity for efficient data processing. In this thesis we take a closer look at both types of these problems. In particular, in Chapter 5 we present an application where we reconstruct images from highly undersampled data by exploiting that images have an intrinsically low complexity. In Chapter 6 we then analyse the theoretical performance of two specific sparse approximation algorithms in a more general setting compared to existing results. For that, in the following, we describe the sparse approximation problem in more detail and introduce the main algorithms which we are going to use.

2.1 Sparse Approximation

In sparse approximation we want to approximate a given signal $y \in \mathbb{R}^d$ by a linear combination of only a small number $S \ll d$ of elements $\phi_i \in \mathbb{R}^d$ out of some given dictionary $\Phi = (\phi_1, \dots, \phi_K)$. This means, we want to find

$$y = \sum_{i \in I} \phi_i x_i + \varepsilon = \Phi_I x_I + \varepsilon \quad \text{such that} \quad |I| = S \ll d \text{ and } \varepsilon \text{ small.} \tag{2.3}$$

The problem of finding the best S-sparse approximation of y in Φ, meaning the best S-support I and coefficient vector x, however is combinatorial. In particular, finding the smallest error for the problem in (2.3) formulates an optimisation problem which is generally NP-hard unless the dictionary forms an orthonormal system. Therefore, in order to solve such problems, suboptimal routines have to be used. By now there exists a large number of such sparse approximation algorithms, where the most popular ones are for example Thresholding, (Orthogonal) Matching Pursuit ((O)MP) [44, 50], Basis Pursuit (BP) [18], or Hard Thresholding Pursuit (HTP) [22]. There exists also detailed theory describing under which conditions they can find the sparse support for $\varepsilon = 0$ or most of the support if ε is small, see e.g. [62, 65, 60, 24, 66]. However, most of these results are only valid under the assumption that we have the signal generating dictionary Φ. Indeed, such assumptions may not always hold, as for instance in the special situation within dictionary learning we are interested in. For that, in Chapter 6 we provide recovery conditions for OMP in situations where we do not have the signal generating dictionary but only a perturbed version of it. We also compare the theoretical and practical performance of OMP with that of the computationally much lighter thresholding algorithm. The insights gained there will also enable us to better understand the practical performance of other methods.

But first, let us give a short reminder of the algorithms we are going to use.

Thresholding. Given a fixed sparsity level S and a dictionary Φ, thresholding finds the S atoms out of the dictionary which are most correlated with the signal and projects it onto their span. In particular, for a given signal y, we

$$\begin{aligned} &\text{find} \quad J = \arg \max_{I:|I|=S} \|\Phi_I^\star y\|_1 \quad \text{and} \\ &\text{calculate} \quad x_J = \Phi_J^\dagger y \quad \text{as well as} \quad \tilde{y} = \Phi_J x_J. \end{aligned}$$

Orthogonal Matching Pursuit (OMP). In each iteration, OMP adds the index corresponding to the atom out of some given dictionary Φ which yields the largest inner product with the current residual. Projecting the signal onto the span of already selected atoms and calculating the new residual, this procedure is repeated until a stopping criterion is met. In particular, initialising with $r_{J_0} = y$ and $J_0 = \emptyset$, we

$$\begin{aligned} &\text{find} \quad j = \arg \max_k |\langle \phi_k, r_{J_i} \rangle| \quad \text{and} \\ &\text{update} \quad J_{i+1} = J_i \cup \{j\} \quad \text{resp.} \quad r_{J_{i+1}} = y - P(\Phi_{J_{i+1}})y. \end{aligned}$$

For many signal classes there exist good predefined dictionaries in which they are sparse, as for example wavelets [15], curvelets [10], or the DCT dictionary which consists of the elements of the discrete cosine transform. However, as such predefined dictionaries do not exist for all signal classes, research has been started into the direction of automatically learning dictionaries providing sparse representations [53, 59]. Further, signal processing tasks as mentioned above can be even more efficiently performed if we have dictionaries yielding very sparse representations with $S \ll d$. This however strongly depends on the good fit between the class of signals and the dictionary. For that, in the following, we describe the concept of dictionary learning in more detail.

2.2 Dictionary Learning

Given a class of signals $y_n \in \mathbb{R}^d$ which are stored in a matrix $Y = (y_1, \dots, y_N) \in \mathbb{R}^{d \times K}$, we want to find a dictionary matrix $\Phi = (\phi_1, \dots, \phi_K) \in \mathbb{R}^{d \times K}$, where each column is normalised, $\|\phi_k\|_2 = 1$, and a sparse coefficient matrix $X = (x_1, \dots, x_N) \in \mathbb{R}^{K \times N}$, such that

$$Y \approx \Phi X \quad \text{with} \quad X \text{ sparse.} \tag{2.4}$$

One way to concretise the abstract formulation in (2.4) is to formulate it as an optimisation problem. For example, given a sparsity level S and a dictionary size K and defining $\mathcal{X}_S$ as the set of all columnwise S-sparse coefficient matrices and $\mathcal{D}_K$ as the set of all dictionaries with K atoms, for some $p \geq 1$, we want to find

$$\operatorname*{argmin}_{\Psi \in \mathcal{D}_K, X \in \mathcal{X}_S} \sum_n \|y_n - \Psi x_n\|_2^p. \tag{2.5}$$

Problems like this are however highly non-convex and hence, very difficult to solve. In order to solve them approximately, one can choose from a wide range of dictionary learning algorithms, e.g. [3, 21, 33, 36, 43, 55]. The dictionary learning algorithms most used in practice belong to the class of alternating optimisation algorithms. This means that they alternate between (trying to) find the best dictionary Ψ while fixing the coefficients X, and the best coefficients X based on the current dictionary Ψ. For example, randomly initialised alternating projection algorithms like K-SVD (K Singular Value Decomposition) for $p = 2$, [3], and ITKrM (Iterative Thresholding and K residual Means) for $p = 1$, [61], tend to be very successful on synthetic data and to provide useful dictionaries on image data. While being computationally very efficient, for these algorithms there exists almost no (K-SVD) or only comparatively weak (ITKrM) theoretical results ensuring dictionary recovery [57, 61]. This is in sharp contrast to more involved graph clustering algorithms and tensor methods which have global recovery guarantees but due to their computational complexity can at best be used in small toy examples, [5, 2, 6]. However, similar results have not yet been shown for learning overcomplete dictionaries via alternating projection algorithms.

Another difficulty which comes along with all dictionary learning algorithms is how to choose the sparsity level S and the dictionary size K. This is quite challenging as their choice can have a large impact on the obtained results and computation time. In general, they are chosen empirically or experimentally as for example in image restoration one will usually find $d \leq K \leq 4d$ and $S = \sqrt{d}$. However, this will probably not always be the best choice.

In this thesis we aim to address some of these problems. In particular, in Chapter 3 we show that ITKrM contracts towards the generating dictionary under much more relaxed conditions compared to those in [61]. Based on an analysis of the convergence behaviour of ITKrM outside the areas where it is a contraction, in Chapter 4, we develop a replacement strategy which finally leads us to a version of ITKrM that adapts both the sparsity level S and the dictionary size K in each iteration. In Chapter 5 we investigate the application of this algorithm to the reconstruction of MR images and compare the results obtained with those of K-SVD and ITKrM. For that purpose, we briefly introduce the latter two in the following.

2.2.1 ITKrM and K-SVD

The ITKrM algorithm was introduced in [61] as a modification of its much simpler predecessor ITKsM (Iterative Thresholding and K signal Means), which uses signal means instead of residual means. From the summary in Algorithm 2.2.1 we can see that ITKrM alternates between updating the sparse support based on the current version of the dictionary using thresholding, and updating the dictionary based on the

Algorithm 2.2.1: ITKrM (one iteration)

```
Input: $\Psi, Y, S$ ;                                    // dictionary, signals, sparsity
Initialise: $\bar{\Psi} = 0$ ;
foreach $n$ do
    $I_n^t = \arg\max_{I:|I|=S} \|\Psi_I^\star y_n\|_1$ ;                       // thresholding
    $a_n = y_n - P(\Psi_{I_n^t})y_n$ ;                                         // residual
    foreach $k \in I_n^t$ do
        $\bar{\psi}_k \leftarrow \bar{\psi}_k + \big[a_n + P(\psi_k)y_n\big] \cdot \mathrm{sign}(\langle \psi_k, y_n \rangle)$ ;     // atom update
    end
end
$\Psi \leftarrow \big(\bar{\psi}_1/\|\bar{\psi}_1\|_2, \dots, \bar{\psi}_K/\|\bar{\psi}_K\|_2\big)$ ;                  // atom normalisation
Output: $\Psi$
```

current support by summing up residuals[1]. Moreover, the signals can be processed sequentially, thus making the algorithm suitable for an online version and parallelisation. Compared to other dictionary learning algorithms, ITKrM exhibits a relatively low computational complexity. Concretely, the determining factors are the matrix vector products $\Psi^\star y_n$ between the current estimate of the dictionary Ψ and the signals, $O(dKN)$, and the projections $P(\Psi_{I_n^t})y_n$. If computed with maximal numerical stability these would have an overall cost $O(S^2dN)$, corresponding to the QR decompositions of $\Psi_{I_n^t}$.

One of the probably most popular and widely used dictionary learning algorithms is K-SVD, which was introduced in [3] as a generalisation of the K-means clustering process. From the summary in Algorithm 2.2.2 we can see that in contrast to ITKrM, K-SVD alternates between updating the sparse support using any sparse approximation algorithm such as OMP, and updating the dictionary by calculating K singular value decompositions instead of calculating K residual means. Compared to ITKrM, K-SVD is computationally more expensive. While K-SVD requires the calculation of K singular value decompositions, its higher computational complexity is also due to the sparse approximation step. K-SVD usually uses OMP to update the sparse support whereas ITKrM uses only simple thresholding.
However, we will not go into further detail here but discuss their performance in Chapter 5. In the next chapter, we will analyse situations in which ITKrM does resp. does not recover the generating dictionary.

[1] In case of ITKsM the atom update rule in Algorithm 2.2.1 is replaced by $\bar{\psi}_k \leftarrow \bar{\psi}_k + y_n \cdot \mathrm{sign}(\langle \psi_k, y_n \rangle)$.

Algorithm 2.2.2: *K*-SVD (one iteration)

```
Input: $\Psi, Y, S$ ;                                        // dictionary, signals, sparsity
Initialise: $R = 0$ ;
foreach $n$ do
    $I_n = \arg\max_{I:|I|=S} \|\Psi_I \Psi_I^\dagger y_n\|_2$ ;                              // via OMP
    $x_n(I_n) = \Psi_{I_n}^\dagger y_n$
    foreach $k \in I_n$ do
        $R_k \leftarrow R_k + \big[y_n - P(\Psi_{I_n})y_n + \psi_k x_n(k)\big]\big[y_n - P(\Psi_{I_n})y_n + \psi_k x_n(k)\big]^\star$
    end
end
foreach $k$ do
    $\bar{\psi}_k = \arg\max_{\|v\|_2=1} \|R_k v\|_2$ ;                                     // via SVD
    $\psi_k \leftarrow \bar{\psi}_k$ ;                                                   // atom update
end
Output: $\Psi$
```

Chapter 3

Contraction Conditions for ITKrM

In this chapter we investigate the contractive behaviour of the Iterative Thresholding and K residual Means (ITKrM) algorithm. After introducing and discussing some existing results, we provide conditions ensuring that one iteration of ITKrM is a contraction under much more relaxed conditions than established previously. The results presented in this chapter are part of some larger work [49].

3.1 Existing Results

We start with a short discussion of some existing results and their shortcomings. In Section 2.2, the ITKrM algorithm was introduced as a modification of its much simpler predecessor ITKsM. One of the main advantages of both algorithms is that they exhibit a relatively low computational complexity. For both algorithms there exists also theory ensuring that they converge locally to a generating dictionary. In particular, for ITKrM it has been shown that if the data is homogeneously S-sparse in a dictionary Φ, where $S \lessapprox \mu^{-2}$, and we initialise with a dictionary Ψ within radius $O(1/\sqrt{S})$, $d(\Psi, \Phi) \lessapprox 1/\sqrt{S}$, then ITKrM using $N = O(K \log K)$ samples in each iteration will converge to the generating dictionary, [61]. For ITKsM a similar result has been proven however, for an even larger convergence radius of size $O(1/\sqrt{\log K})$.
Comparing the theoretical results for ITKrM with its practical performance, in simulations on synthetic data it shows even better convergence behaviour. Concretely, if the atoms of the generating dictionary are perturbed with vectors z_k chosen uniformly at random from the sphere, $\psi_k = \alpha_k \phi_k + \omega_k z_k$, ITKrM converges also for ratios $\alpha_k : \omega_k = 1 : 4$. For completely random initialisations, $\psi_k = z_k$, it finds between 90% and 100% of the atoms - depending on the noise and sparsity level. Also on image data ITKrM produces dictionaries of the same quality as K-SVD in a fraction of the time, [47]. For ITKsM on the other hand it has been shown that in case of the $1:4$ initialisations, its recovery rate deteriorates quite drastically as the sparsity level S increases. In case of random initialisations, the recovery rates of ITKsM are at best around 70% however only for noiseless signals and very small sparsity levels. In the noisy setting this recovery rate further decreases down to around 35% for increasing S, [61].

Considering the good practical performance of ITKrM, it is especially frustrating that we only get a convergence radius of size $O(1/\sqrt{S})$, while for its simpler predecessor ITKsM, which when initialised randomly performs much worse both on synthetic and image data, one can prove a convergence radius of size $O(1/\sqrt{\log K})$. For that, in the following, we will take a closer look at the two algorithms and the differences in the convergence proofs. This will allow us to show that ITKrM behaves well on a much larger area.

Differences in convergence proofs

To better understand the idea behind the convergence proofs we first rewrite the atom update formula before normalisation, which for one iteration of ITKrM becomes

$$\bar{\psi}_k = \sum_{n: k \in I_n^t} \left[\mathbb{I}_d - P(\Psi_{I_n^t}) + P(\psi_k)\right] y_n \cdot \operatorname{sign}(\langle \psi_k, y_n \rangle),$$

while for ITKsM we can take the formula above and simply ignore the operators in the square brackets. Adding and replacing some terms, we expand the sum as

$$
\begin{aligned}
\bar{\psi}_k \quad = \quad & \left.\begin{aligned}
&\sum_{n:k\in I_n^t} \big[\mathbb{I}_d - P(\Psi_{I_n^t}) + P(\psi_k)\big] y_n \cdot \operatorname{sign}(\langle \psi_k, y_n\rangle) \\
&\qquad\qquad - \sum_{n:k\in I_n} \big[\mathbb{I}_d - P(\Psi_{I_n}) + P(\psi_k)\big] y_n \cdot \sigma_n(k)
\end{aligned}\right\} S_1 \\
& \left.\begin{aligned}
+ &\sum_{n:k\in I_n} \big[\mathbb{I}_d - P(\Psi_{I_n}) + P(\psi_k)\big] y_n \cdot \sigma_n(k) \\
&\qquad\qquad - \sum_{n:k\in I_n} \big[\mathbb{I}_d - P(\Phi_{I_n}) + P(\phi_k)\big] y_n \cdot \sigma_n(k)
\end{aligned}\right\} S_2 \\
& \left. + \sum_{n:k\in I_n} \big[y_n - P(\Phi_{I_n}) y_n + P(\phi_k) y_n\big] \cdot \sigma_n(k). \right\} S_3
\end{aligned}
$$

The term S_1 captures the errors which thresholding makes in estimating the supports I_n and signs $\sigma_n(k) = \operatorname{sign}(\langle \phi_k, y_n\rangle)$ when using the current estimate Ψ. We know that it is (sufficiently) small as long $d(\Phi, \Psi) \lesssim 1/\sqrt{\log K}$. The second term S_2 captures the difference between the residual using the current estimate and the true dictionary, respectively, which is small as long as $d(\Phi, \Psi) \lesssim 1/\sqrt{S}$. In expectation the last term is simply a multiple of the true atom ϕ_k. Hence, as long as the number of signals N is large enough, it will concentrate arbitrarily close to ϕ_k.

From this we can see that the main constraint on the convergence radius for ITKrM stems from the second term S_2, which simply vanishes in case of ITKsM. The problem is that we need to invert the $S \times S$ matrix $\Psi_{I_n}^\star \Psi_{I_n}$, which is a perturbed version of the matrix $\Phi_{I_n}^\star \Phi_{I_n}$. If the difference between the dictionaries scales as $d(\Phi, \Psi) \approx 1/\sqrt{S}$, then there exist perturbations such that $\Psi_{I_n}^\star \Psi_{I_n}$ is ill conditioned even if $\Phi_{I_n}^\star \Phi_{I_n}$ is not. However, from [66, 14] we know that if the current dictionary estimate Ψ itself is a well-conditioned and incoherent matrix, for most possible supports I_n, $\Psi_{I_n}^\star \Psi_{I_n}$ will be close to the identity as long as $S \lesssim d/\log K$. This means that the term S_2 should be small as long as the current estimate Ψ is well-conditioned and incoherent, a property which can be verified after each iteration.

Therefore, the next question is if also the first term S_1 can be controlled for a larger class of dictionaries Ψ. In previous estimates this error was bounded for each atom by the probability of thresholding failing multiplied with the norm bound on the difference of the projections. While this strategy is simple, it is quite crude as it assigns any error of thresholding to all atoms. However, an atom $\bar{\psi}_k$ is only affected by a thresholding error if either k is in the original support or if k is not in the original support but is included in the thresholded support. Further, we can take into account that by perturbing an atom ϕ_k, meaning $\psi_k = \alpha_k \phi_k + \omega_k z_k$, its coherence to one other atom ϕ_ℓ may increase dramatically - to the point of it being a better approximant than ψ_ℓ, that is, if $z_k \approx \phi_\ell$ we get $\langle \phi_k, \phi_\ell\rangle \ll \langle \psi_k, \phi_\ell\rangle \approx \langle \psi_\ell, \phi_\ell\rangle$. However, if the original Φ itself is well-conditioned, ψ_k cannot become coherent to all (many) other atoms.

Indeed, using both of these ideas we get a refined result characterising the contractive areas of ITKrM. Before presenting our result in the next section, in the following, we introduce the signal model on which all our theoretical findings are based.

3.1.1 Sparse signal model

As basis for our results we use the following signal model, already used in [57, 58, 61]. Given a $d \times K$ dictionary Φ, we assume that the signals are generated as

$$y = \frac{\Phi x + r}{\sqrt{1 + \|r\|_2^2}}, \tag{3.1}$$

where $x \in \mathbb{R}^K$ is a sparse coefficient sequence and $r \in \mathbb{R}^d$ is some noise. We assume that r is a centered sub-Gaussian vector with parameter ρ, that is, $\mathbb{E}(r) = 0$ and for all vectors v the marginals $\langle v, r \rangle$ are sub-Gaussian with parameter ρ, meaning they satisfy $\mathbb{E}(e^{t\langle v,r\rangle}) \leq e^{t^2\rho^2/2}$ for all $t > 0$.
To model the coefficient sequences x we first assume that there is a measure ν_c on a subset $\mathcal{C}$ of the positive, non increasing sequences with unit norm, meaning for $c \in \mathcal{C}$ we have $c(1) \geq c(2) \ldots \geq c(K) \geq 0$ and $\|c\|_2 = 1$. A coefficient sequence x is created by drawing a sequence c according to ν_c, and both a permutation p and a sign sequence σ uniformly at random and setting $x = x_{c,p,\sigma}$, where $x_{c,p,\sigma}(k) = \sigma(k)c(p(k))$. The signal model then takes the form

$$y = \frac{\Phi x_{c,p,\sigma} + r}{\sqrt{1 + \|r\|_2^2}}. \tag{3.2}$$

Using this model it is quite simple to incorporate sparsity via the measure ν_c. To model approximately S-sparse signals we require that the S largest absolute coefficients, meaning those inside the support $I = p^{-1}(\mathbb{S})$, are well balanced and much larger than the remaining ones outside the support. Further, we need that the expected energy of the coefficients outside the support is relatively small and that the sparse coefficients are well separated from the noise. Concretely we require that almost ν_c-surely we have

$$\frac{c(1)}{c(S)} \leq \gamma_{dyn}, \qquad \frac{c(S+1)}{c(S)} \leq \gamma_{gap}, \qquad \frac{\|c(\mathbb{S}^c)\|_2}{c(1)} \leq \gamma_{app} \qquad \text{and} \qquad \frac{\rho}{c(S)} \leq \gamma_\rho. \tag{3.3}$$

We will refer to the worst case ratio between coefficients inside the support, γ_{dyn}, as dynamic (sparse) range and to the worst case ratio between coefficients outside the support to those inside the support, γ_{gap}, as the (sparse) gap. Since for a noise free signal the expected squared sparse approximation error is

$$\mathbb{E}(\|\sum_{k \notin I} \sigma_k c(p(k)) \phi_k\|_2^2) = \|c(\mathbb{S}^c)\|_2^2,$$

we will call γ_{app} the relative (sparse) approximation error. Finally, γ_ρ is called the noise to (sparse) coefficient ratio.
Apart from these worst case bounds we will also use three other signal statistics,

$$\gamma_{1,S} := \mathbb{E}_c\left(\|c(\mathbb{S})\|_1\right), \qquad \gamma_{2,S} := \mathbb{E}_c\left(\|c(\mathbb{S})\|_2^2\right), \qquad C_r := \mathbb{E}_r\left(\frac{1}{\sqrt{1+\|r\|_2^2}}\right). \tag{3.4}$$

The constant $\gamma_{1,S}$ helps to characterise the average size of the sparse coefficients, $\gamma_{1,S} = \mathbb{E}(|x_i| : i \in I) \cdot S \leq \sqrt{S}$, while $\gamma_{2,S}$ characterises the average sparse approximation quality, $\gamma_{2,S} = \mathbb{E}(\|\Phi_I x_I\|_2^2) \leq 1$. The noise constant can be bounded by

$$C_r \geq \frac{1-e^{-d}}{\sqrt{1+5d\rho^2}}, \tag{3.5}$$

and for large ρ approaches the signal to noise ratio, $C_r^2 \approx \frac{1}{d\rho^2} \approx \frac{\mathbb{E}(\|\Phi x\|_2^2)}{\mathbb{E}(\|r\|_2^2)}$, see [58] for details.
To get a better feeling for all the involved constants, we will calculate them for the case of perfectly sparse signals where $c(i) = 1/\sqrt{S}$ for $i \leq S$ and $c(i) = 0$ else. We then have $\gamma_{dyn} = 1$, $\gamma_{gap} = 0$ and $\gamma_{app} = 0$ as well as $\gamma_{1,S} = \sqrt{S}$ and $\gamma_{2,S} = 1$. In the case of noiseless signals we have $C_r = 1$ and $\gamma_\rho = 0$. In the case of Gaussian noise the noise to coefficient ratio is related to the signal to noise ratio via $\mathrm{SNR} = S/(\gamma_\rho^2 d)$.

3.2 Contraction Theorem

Here we state and prove our refined contraction theorem. Note that, we only consider distances $d(\Psi,\Phi) \geq \frac{1}{32\sqrt{S}}$. The result for the case $d(\Psi,\Phi) \leq \frac{1}{32\sqrt{S}}$ can be found in [61].

Theorem 3.1. *Assume that the signals y_n follow model* (3.2) *for a dictionary Φ with $\|\Phi\|_{2,2}^2 \leq \frac{K}{98S}$ and for coefficients with gap $c(S+1)/c(S) \leq \gamma_{gap}$, dynamic sparse range $c(1)/c(S) \leq \gamma_{dyn}$, noise to coefficient ratio $\rho/c(S) \leq \gamma_\rho$ and relative approximation error $\|c(\mathbb{S}^c)\|_2/c(1) \leq \gamma_{app} \leq \frac{12}{7}\sqrt{\log K}$. Further, assume that the coherence and operator norm of the current dictionary estimate Ψ satisfy,*

$$\mu(\Psi) \leq \frac{1}{20\log K} \qquad \textit{and} \qquad \|\Psi\|_{2,2}^2 \leq \frac{K}{134e^2 S\log K} - 1. \tag{3.6}$$

If $d(\Psi,\Phi) \geq \frac{1}{32\sqrt{S}}$ but the cross-Gram matrix $\Phi^\star\Psi$ is diagonally dominant in the sense *that*

$$\min_k |\langle \psi_k, \phi_k\rangle| \geq \max\Bigg\{8\,\gamma_{gap}\cdot\max_k |\langle\psi_k,\phi_k\rangle|,\; 40\,\gamma_\rho\cdot\sqrt{\log K},\; 48\,\gamma_{dyn}\cdot\log K\cdot\mu(\Phi,\Psi),\; 14\,\gamma_{dyn}\cdot\sqrt{\|\Phi\|_{2,2}^2 S\log K/(K-S)}\Bigg\}, \tag{3.7}$$

then one iteration of ITKrM using N training signals will reduce the distance by at least a factor $\kappa \leq 0.95$, meaning $d(\bar{\Psi},\Phi) \leq 0.95\cdot d(\Psi,\Phi)$, except with probability

$$3K\exp\left(-\frac{NC_r^2\gamma_{1,S}^2\cdot\varepsilon}{768K\max\{S,\|\Phi\|_{2,2}^2+1\}^{\frac{3}{2}}}\right) + 4K\exp\left(-\frac{NC_r^2\gamma_{1,S}^2\cdot\varepsilon^2}{512K\max\{S,\|\Phi\|_{2,2}^2+1\}\,(1+d\rho^2)}\right).$$

Before we prove the theorem we would like to say a few words about the result. The conditions in (3.6) simply say that we have to exclude dictionaries Ψ which are coherent or have large operator norm. From [61] we know that ITKrM succeeds if the input dictionary is within a ball of radius $1/(32\sqrt{S})$ around the generating dictionary Φ. If we are in an area outside this ball, Theorem 3.1 says that ITKrM is a contraction towards Φ whenever the additional condition in (3.7) is satisfied. Taking a closer look at the condition on the cross-Gram matrix, the determining factors are

$$48\,\gamma_{dyn}\cdot\log K\cdot\mu(\Phi,\Psi) \quad\text{and}\quad 14\,\gamma_{dyn}\cdot\|\Phi\|_{2,2}\sqrt{S\log K/(K-S)}.$$

In particular, the fact that the diagonal entries have to be larger than $14\,\gamma_{dyn}\cdot\|\Phi\|_{2,2}\sqrt{S\log K/(K-S)}$ puts a constraint on the admissible distance $d(\Phi,\Psi)$ via the relation $d(\Phi,\Psi)^2 = 2-2\min_k|\langle\phi_k,\psi_k\rangle|$. For example, for a well-conditioned dictionary satisfying $\|\Phi\|_{2,2}^2 \approx K/d$, this means that

$$d(\Phi,\Psi) \lesssim \left(2-2\sqrt{\frac{S\log K}{d}}\right)^{1/2}. \tag{3.8}$$

Considering that the maximal distance between two dictionaries is $\sqrt{2}$, this is a large improvement over the admissible distance $1/(32\sqrt{S})$ in previous results. However, the additional price to pay is that also the intrinsic condition on the cross-Gram matrix needs to be satisfied, meaning,

$$\min_k|\langle\psi_k,\phi_k\rangle| \geq 48\,\gamma_{dyn}\cdot\log K\cdot\max_{j\neq k}|\langle\phi_k,\psi_j\rangle|. \tag{3.9}$$

This condition captures our intuition that two estimated atoms should not point to the same generating atom and provides a bound for sufficient separation.
Note that Theorem 3.1 does not guarantee convergence of ITKrM since it is only valid for one iteration. In order to prove convergence we would additionally have to prove that $\bar{\Psi}$ inherits from Ψ the properties that are required for being a contraction, which, however, is part of our future goals. Nevertheless, the result contributes significantly to explaining the good convergence behaviour of ITKrM.
For example, it allows us to briefly sketch why the algorithm always converges in experiments where the initial dictionary is a large but random perturbation of a well-behaved generating dictionary Φ with coherence $\mu(\Phi) \approx 1/\sqrt{d}$ and operatornorm $\|\Phi\|_{2,2}^2 \approx K/d$. For $\psi_k = \alpha_k \phi_k + \omega_k z_k$, where the perturbation vectors z_k are drawn uniformly at random from the unit sphere orthogonal to ϕ_k, with high probability, we have for all $j \neq k$

$$|\langle \phi_k, z_j \rangle| \lesssim \sqrt{\log K/d} \quad \text{and} \quad |\langle z_k, z_j \rangle| \lesssim \sqrt{\log K/d}, \tag{3.10}$$

and consequently for all possible α_k

$$\mu(\Psi) \lesssim \sqrt{4 \log K/d} \quad \text{and} \quad \mu(\Phi, \Psi) \lesssim \sqrt{2 \log K/d}. \tag{3.11}$$

Also with high probability the operator norm of the matrix $Z = (z_1, \dots z_K)$ is bounded by $\|Z\|_{2,2} \lesssim \sqrt{\log K}$, [64], so that for Ψ we get $\|\Psi\|_{2,2} \lesssim \sqrt{K/d} + \sqrt{\log K}$, again independent of α_k. Comparing these estimates with the requirements of the theorem we see that for moderate sparsity levels, $S \geq \log K$, we get a contraction whenever

$$\alpha_{\min} \gtrsim \sqrt{\frac{S(\log K)^2}{d}} \qquad \Leftrightarrow \qquad d(\Phi, \Psi) \lesssim \left(2 - 2\sqrt{\frac{S(\log K)^2}{d}}\right)^{1/2}. \tag{3.12}$$

In the following, we state the proof of Theorem 3.1.

Proof. We follow the outline of the proof for Theorem 4.2 in [61]. However, to extend the convergence radius we need to introduce new ideas. First for bounding the probability of thresholding with Ψ not recovering the generating support or preserving the generating sign, replacing Lemma B.3/4 of [61], and second for bounding the difference between the oracle residuals based on Ψ and Φ, replacing Lemma B.8 of [61]. In order to make the ideas precise let us introduce the following definitions. We denote the thresholding residual based on Ψ by

$$R^t(\Psi, y_n, k) := \left[y_n - P(\Psi_{I^t_{\Psi,n}})y_n + P(\psi_k)y_n\right] \cdot \operatorname{sign}(\langle \psi_k, y_n \rangle) \cdot \chi(I^t_{\Psi,n}, k), \tag{3.13}$$

and the oracle residual based on the generating support $I_n = p_n^{-1}(\mathbb{S})$, the generating signs σ_n and Ψ, by

$$R^o(\Psi, y_n, k) := \left[y_n - P(\Psi_{I_n})y_n + P(\psi_k)y_n\right] \cdot \sigma_n(k) \cdot \chi(I_n, k). \tag{3.14}$$

Hence, we can write

$$\begin{aligned}\bar{\psi}_k &= \frac{1}{N}\sum_n \left[R^t(\Psi,y_n,k) - R^o(\Psi,y_n,k)\right] + \frac{1}{N}\sum_n \left[R^o(\Psi,y_n,k) - R^o(\Phi,y_n,k)\right]\\ &\qquad + \frac{1}{N}\sum_n R^o(\Phi,y_n,k)\\ &= \frac{1}{N}\sum_n \left[R^t(\Psi,y_n,k) - R^o(\Psi,y_n,k)\right] + \frac{1}{N}\sum_n \left[R^o(\Psi,y_n,k) - R^o(\Phi,y_n,k)\right]\\ &\qquad + \frac{1}{N}\sum_n \left[y_n - P(\Phi_{I_n})y_n\right]\cdot\sigma_n(k)\cdot\chi(I_n,k)\\ &\qquad + \left(\frac{1}{N}\sum_n \langle y_n,\phi_k\rangle\cdot\sigma_n(k)\cdot\chi(I_n,k)\right)\phi_k. \qquad (3.15)\end{aligned}$$

Using the abbreviation $s_k = \frac{1}{N}\sum_n \langle y_n,\phi_k\rangle\cdot\sigma_n(k)\cdot\chi(I_n,k)$, we obtain

$$\begin{aligned}\|\bar{\psi}_k - s_k\phi_k\|_2 &\le \frac{1}{N}\Big\|\sum_n \left[R^t(\Psi,y_n,k) - R^o(\Psi,y_n,k)\right]\Big\|_2\\ &\qquad + \frac{1}{N}\Big\|\sum_n \left[R^o(\Psi,y_n,k) - R^o(\Phi,y_n,k)\right]\Big\|_2\\ &\qquad + \frac{1}{N}\Big\|\sum_n \left[y_n - P(\Phi_{I_n})y_n\right]\cdot\sigma_n(k)\cdot\chi(I_n,k)\Big\|_2. \qquad (3.16)\end{aligned}$$

The first norm term gives the error originating from thresholding failing to recover the generating support I_n and/or preserving the generating sign σ_n. Assuming that the generating support I_n is recovered, the second norm term gives the difference of the residuals using Φ and Ψ, respectively. The last norm term covers the residual energy when projecting onto the orthogonal complement of Φ_{I_n}, meaning, the signal energy that remains when projecting onto the subspace spanned by the atoms indexed by $i \in I_n$.

In the following we show that these terms are small with high probability and hence, that one iteration of ITKrM reduces the error by at least a factor $\kappa < 1$. In particular, setting $B = \|\Phi\|_{2,2}^2$ and $\varepsilon = d(\Psi,\Phi)$, for the first norm term on the right hand side of (3.16), by Lemma 3.5 in Subsection 3.2.1, we have

$$\begin{aligned}\mathbb{P}&\left(\frac{1}{N}\Big\|\sum_n \left[R^t(\Psi,y_n,k) - R^o(\Psi,y_n,k)\right]\Big\|_2 > \frac{18(S+1)\sqrt{B+1}}{K^3} + \frac{C_r\gamma_{1,S}}{K}t_1\varepsilon\right)\\ &\qquad \le 2\exp\left(-\frac{NC_r^2\gamma_{1,S}^2t_1^2\varepsilon^2}{\frac{108(S+1)(B+1)}{K} + 3t_1\varepsilon C_r\gamma_{1,S}K\sqrt{B+1}}\right). \qquad (3.17)\end{aligned}$$

For the second norm term, by Lemma 3.7 in Subsection 3.2.2, we have that for $0 \leq t_2 \leq 1/8$ and $S \leq \min\left\{\frac{K}{98B}, \frac{1}{98\rho^2}\right\}$,

$$\mathbb{P}\left(\frac{1}{N}\Big\|\sum_n [R^o(\Psi, y_n, k) - R^o(\Phi, y_n, k)]\Big\|_2 \geq \frac{C_r\gamma_{1,S}}{K}(0.308\varepsilon + t_2\varepsilon)\right) \leq \exp\left(-\frac{NC_r^2\gamma_{1,S}^2 \cdot t_2^2\varepsilon}{12K\max\{S,B\}^{\frac{3}{2}}} + \frac{1}{4}\right). \quad (3.18)$$

For the remaining terms we will use the bounds already derived in [61]. In particular, by Lemma B.6 from [61], we have

$$\mathbb{P}\left(\Big|\frac{1}{N}\sum_n \chi(I_n, k)\sigma_n(k)\langle y_n, \phi_k\rangle\Big| \leq (1-t_0)\frac{C_r\gamma_{1,S}}{K}\right) \leq \exp\left(-\frac{NC_r^2\gamma_{1,S}^2 \cdot t_0^2}{2K(1 + \frac{SB}{K} + S\rho^2 + t_0C_r\gamma_{1,S}\sqrt{B+1}/3)}\right), \quad (3.19)$$

and by Lemma B.7 from [61], we have that

$$\mathbb{P}\left(\Big\|\frac{1}{N}\sum_n \left[y_n - P(\Phi_{I_n})y_n\right] \cdot \sigma_n(k) \cdot \chi(I_n, k)\Big\|_2 \geq \frac{C_r\gamma_{1,S}}{K}t_3\varepsilon\right) \leq \exp\left(-\frac{NC_r^2\gamma_{1,S}^2 \cdot t_3\varepsilon}{8K\max\{S, B+1\}} \cdot \min\left\{\frac{t_3\varepsilon}{(1-\gamma_{2,S}+d\rho^2)}, 1\right\} + \frac{1}{4}\right). \quad (3.20)$$

Putting all these pieces together, with high probability, we have $s_k \geq (1-t_0)\frac{C_r\gamma_{1,S}}{K}$ and

$$\|\bar{\psi}_k - s_k\phi_k\|_2 \leq \frac{C_r\gamma_{1,S}}{K}\left(\frac{18(S+1)\sqrt{B+1}}{K^2C_r\gamma_{1,S}\varepsilon} + t_1 + 0.308 + t_2 + t_3\right)\varepsilon. \quad (3.21)$$

Note that we only need to take into account distances $\varepsilon \geq \frac{1}{32\sqrt{S}}$, so we will use some crude bounds on $C_r\gamma_{1,S}$ to show that the fraction with ε in the denominator above is small. The requirement that $\|c(\mathbb{S}^c)\|_2/c(1) \leq \gamma_{app} \leq \frac{12}{7}\sqrt{\log K}$ ensures that $\gamma_{1,S} \geq (1 + 3\log K)^{-1/2}$ and we trivially have $\gamma_{1,S} \geq Sc(S)$. In particular, we have that

$$\gamma_{1,S} = \mathbb{E}_c(\|c(\mathbb{S})\|_1) = \mathbb{E}_c(c(1) + \cdots + c(S)) \geq Sc(S),$$

and as the coefficient sequences c are normalised, we have

$$\|c(\mathbb{S})\|_2^2 \geq 1 - \|c(\mathbb{S}^c)\|_2^2 \geq 1 - c(1)^2 3\log K,$$

which in turn means

$$\|c(\mathbb{S})\|_1 \geq \|c(\mathbb{S})\|_2 \geq \frac{1}{\sqrt{1+3\log K}}.$$

Combining this with the bound on C_r in (3.5),

$$C_r \geq \frac{1-e^{-d}}{\sqrt{1+5d\rho^2}},$$

we get

$$\frac{1}{C_r\gamma_{1,S}} \leq \frac{\sqrt{1+5d\rho^2}}{(1-e^{-d})\gamma_{1,S}} \leq \frac{\sqrt{1+3\log K}}{(1-e^{-d})} + \frac{\rho}{c(S)}\frac{\sqrt{5d}}{S(1-e^{-d})}. \tag{3.22}$$

The conditions in (3.7) imply that $K \geq 14^2 SB\log K$, which in turn means that $\log K > 7$, as well as $\rho/c(S) \leq \gamma_\rho \leq 1/(40\sqrt{\log K})$. Assuming additionally that $K \geq \sqrt{d}$, meaning the dictionary is not too undercomplete, this leads to

$$\begin{aligned}
\frac{18\cdot(S+1)\sqrt{B+1}}{K^2C_r\gamma_{1,S}\varepsilon} &\leq \frac{18\cdot 32\cdot(S+1)\sqrt{S(B+1)}}{K^2C_r\gamma_{1,S}}\\
&\leq \frac{18\cdot 32\cdot(S+1)\sqrt{S(B+1)}}{K^2}\left(\frac{\sqrt{1+3\log K}}{1-e^{-d}} + \frac{\rho}{c(S)}\frac{\sqrt{5d}}{S(1-e^{-d})}\right)\\
&\leq \frac{18\cdot 32\cdot(S+1)\sqrt{S(B+1)}}{K^2}\left(\frac{\sqrt{1+3\log K}}{1-e^{-d}} + \frac{\sqrt{5d}}{40S\sqrt{\log K}(1-e^{-d})}\right)\\
&\leq \frac{8\cdot 18\cdot 32\cdot(S+1)\sqrt{S(B+1)}}{5K^2}\left(\sqrt{1+3\log K} + \frac{\sqrt{5d}}{40S\sqrt{\log K}}\right)\\
&\leq \frac{8\cdot 18\cdot 32}{5}\left(\frac{(S+1)\sqrt{S(B+1)}\sqrt{1+3\log K}}{14^4S^2B^2\log^2 K} + \frac{(S+1)\sqrt{S(B+1)}\sqrt{5}}{40\cdot 14^2S^2B\log K\sqrt{\log K}}\right)\\
&\leq \frac{8\cdot 18\cdot 32}{5}\left(\frac{3}{14^5\cdot\sqrt{2}} + \frac{3\cdot\sqrt{5}}{40\cdot 14^3\cdot\sqrt{7}}\right) \leq 0.025,
\end{aligned}$$

for $B \geq 1$ and $S \geq 2$. Setting $t_0 = t_1 = 1/20$ and $t_2 = t_3 = 1/8$ we get

$$\max_k \left\|\bar{\psi}_k - s_k\phi_k\right\|_2 \leq 0.633\cdot\frac{C_r\gamma_{1,S}}{K}\varepsilon \quad \text{and} \quad \min_k s_k \geq 0.95\cdot\frac{C_r\gamma_{1,S}}{K}, \tag{3.23}$$

which by Lemma B.10 from [61] implies that

$$\begin{aligned}
d(\bar{\Psi},\Phi)^2 = \max_k\left\|\frac{\bar{\psi}_k}{\|\bar{\psi}_k\|_2} - \phi_k\right\|_2^2 &\leq 2\left(1-\sqrt{1-\frac{0.633^2\varepsilon^2}{0.95^2}}\right)\\
&\leq \frac{2\cdot 0.633^2\varepsilon^2}{0.95^2} \leq 0.89\varepsilon^2,
\end{aligned} \tag{3.24}$$

except with probability

$$K\exp\left(-\frac{NC_r^2\gamma_{1,S}^2}{K(801+14C_r\gamma_{1,S}\sqrt{B{+}1})}\right)$$
$$+2K\exp\left(-\frac{NC_r^2\gamma_{1,S}^2\cdot\varepsilon^2}{K(\frac{1}{10}+60\varepsilon C_r\gamma_{1,S}\sqrt{B{+}1})}\right)+e^{\frac{1}{4}}K\exp\left(-\frac{NC_r^2\gamma_{1,S}^2\cdot\varepsilon}{768K\max\{S,B\}^{\frac{3}{2}}}\right)$$
$$+e^{\frac{1}{4}}K\exp\left(-\frac{NC_r^2\gamma_{1,S}^2\cdot\varepsilon^2}{512K\max\{S,B{+}1\}\,(1+d\rho^2)}\right).$$

The final probability bound follows from the observations that $C_r\gamma_{1,S}\leq\sqrt{S}$, $B{+}1\geq 2$ and $\varepsilon\leq\sqrt{2}$. □

In the following two subsections we derive the two lemmata characterising the difference between the thresholding and the oracle residual (Lemma 3.5) resp. the difference between the oracle residuals based on the generating dictionary and a perturbation (Lemma 3.7).

3.2.1 Difference between thresholding and oracle residual

In order to prove Lemma 3.5 we will make use of the scalar version of Bernstein's inequality [8] and Hoeffding's inequality [29]. We will also need Proposition 3.4 to deal with sums of dependent random variables. The proof of Proposition 3.4 can be found in Appendix A.1.

Theorem 3.2 (Scalar Bernstein, [8]). *Let $v_n\in\mathbb{R}$, $n=1\dots N$, be a finite sequence of independent random variables with zero mean. If $\mathbb{E}(v_n^2)\leq m$ and $\mathbb{E}(|v_n|^k)\leq\frac{1}{2}k!\,mM^{k-2}$ for all $k>2$, then for all $t>0$ we have*

$$\mathbb{P}\left(\sum_n v_n\geq t\right)\leq\exp\left(-\frac{t^2}{2(Nm+Mt)}\right).$$

Theorem 3.3 (Hoeffding, [29]). *Let $X_1,\dots,X_n$ be independent random variables and $a_i\leq X_i\leq b_i$ for all i. Then for all $t>0$ we have*

$$\mathbb{P}\left(\frac{1}{n}\Big|\sum_{i=1}^n(X_i-\mathbb{E}(X_i))\Big|\geq t\right)\leq 2\exp\left(-\frac{2n^2t^2}{\sum_{i=1}^n(b_i-a_i)^2}\right).$$

Proposition 3.4. *Let $v\in\mathbb{R}^K$ be a vector, $I=(i_1,\dots,i_S)$ be a sequence of length S obtained by sampling from $\mathbb{K}=\{1,\dots,K\}$ without replacement, ε with values in $\{-1,1\}^S$ a Rademacher vector independent from I and $c\in\mathbb{R}^S$ a scaling vector. Then for any $t\geq 0$,*

$$\mathbb{P}\left(\Big|\sum_{k=1}^S c_k\varepsilon_k v_{i_k}\Big|\geq t\right)\leq 2\exp\left(-\frac{t^2}{2\big(\|c\|_\infty\|v\|_\infty t+\|c\|_2^2\|v\|_2^2/(K-S)\big)}\right). \tag{3.25}$$

Being equipped with the right tools, we are now ready to prove the lemma estimating the error originating from thresholding failing to recover the generating support and signs.

Lemma 3.5. *Assume that the signals y_n follow model* (3.2) *for coefficients with gap $c(S{+}1)/c(S) \leq \gamma_{gap}$, dynamic sparse range $c(1)/c(S) \leq \gamma_{dyn}$, noise to coefficient ratio $\rho/c(S) \leq \gamma_\rho$ and relative approximation error $\|c(\mathbb{S}^c)\|_2/c(1) \leq \gamma_{app} \leq \frac{12}{7}\sqrt{\log K}$. If the cross-Gram matrix $\Phi^\star\Psi$ is diagonally dominant in the sense that*

$$\begin{aligned}\min_k |\langle \psi_k, \phi_k\rangle| \geq \max\Big\{ & 8\,\gamma_{gap} \cdot \max_k |\langle \psi_k, \phi_k\rangle|, \\ & 40\,\gamma_\rho \cdot \sqrt{\log K}, \\ & 48\,\gamma_{dyn} \cdot \log K \cdot \mu(\Phi,\Psi), \\ & 14\,\gamma_{dyn} \cdot \sqrt{\|\Phi\|_{2,2}^2 S \log K/(K{-}S)}\Big\}, \end{aligned} \tag{3.26}$$

then

$$\begin{aligned}\mathbb{P}\left(\frac{1}{N}\Big\|\sum_n \big[R^t(\Psi,y_n,k) - R^o(\Psi,y_n,k)\big]\Big\|_2 > \frac{18(S{+}1)\sqrt{\|\Phi\|_{2,2}^2{+}1}}{K^3} + \frac{C_r\gamma_{1,S}}{K}t\varepsilon\right) \\ \leq 2\exp\left(-\frac{NC_r^2\gamma_{1,S}^2t^2\varepsilon^2}{\frac{108(S+1)(\|\Phi\|_{2,2}^2+1)}{K} + 3t\varepsilon C_r\gamma_{1,S}K\sqrt{\|\Phi\|_{2,2}^2{+}1}}\right).\end{aligned} \tag{3.27}$$

Proof. Throughout the proof we use the abbreviations $B = \|\Phi\|_{2,2}^2$ and $\hat\mu = \mu(\Phi,\Psi)$. To estimate the difference between the oracle and the thresholding residuals, we have to distinguish between four different cases, based on whether k is in the oracle support or not and whether thresholding recovers the oracle support and sign, so we set

$$\begin{aligned}\mathcal{F} &= \{n : k \in I_n \wedge \big(I_n^t \neq I_n \vee \operatorname{sign}(\langle\psi_k, y_n\rangle) \neq \sigma_n(k)\big)\},\\ \mathcal{G} &= \{n : k \notin I_n \wedge k \in I_n^t\}.\end{aligned}$$

Whenever a signal is not in one of the sets above, the residuals coincide, yielding

$$\begin{aligned}\Delta &= \Big\|\sum_n \big[R^t(\Psi,y_n,k) - R^o(\Psi,y_n,k)\big]\Big\|_2 \\ &= \Big\|\sum_{n\in\mathcal{F}\cup\mathcal{G}} \big[R^t(\Psi,y_n,k) - R^o(\Psi,y_n,k)\big]\Big\|_2.\end{aligned} \tag{3.28}$$

Further, observing that operators of the form $\mathbb{I}_d - P(\Psi_J) + P(\psi_k)$ with $k \in J$ are orthogonal projections, and that our signals are bounded, $\|y_n\|_2 \leq \sqrt{B+1}$, as well as $R^o(\Psi,y_n,k) = 0$ for $n \in \mathcal{G}$, leads to

$$\Delta \leq \sum_{n\in\mathcal{F}\cup\mathcal{G}} \big(\|R^t(\Psi,y_n,k)\|_2 + \|R^o(\Psi,y_n,k)\|\big) \leq (2|\mathcal{F}| + |\mathcal{G}|)\sqrt{B+1}. \tag{3.29}$$

To upper bound the size of the set $\mathcal{F}$, we apply Bernstein's inequality to the sum of N i.i.d copies of the centered random variable $\mathbf{1}_F - \mathbb{P}(F)$, where

$$F = \big\{y : k \in I \wedge \big(I^t \neq I \vee \operatorname{sign}(\langle \psi_k, y\rangle) \neq \sigma(k)\big)\big\}\,, \tag{3.30}$$

which leads to

$$\mathbb{P}(|\mathcal{F}| \geq N\mathbb{P}(F) + Nt) \leq \exp\left(-\frac{t^2 N}{2\mathbb{P}(F) + t}\right). \tag{3.31}$$

Similarly defining $G = \{y : k \notin I \wedge k \in I^t\}$, we get

$$\mathbb{P}(|\mathcal{G}| \geq N\mathbb{P}(G) + Nt) \leq \exp\left(-\frac{t^2 N}{2\mathbb{P}(G) + t}\right). \tag{3.32}$$

So what remains to calculate is the probability of the events F and G, that is of thresholding failing to recover the oracle support and sign when k is in the support and of accidentally recovering k when it is not in the support.

Step 1 - Failure probability of the recovery of I or the correct sign $\sigma(k)$

Here we show that with high probability for a signal y following the model in (3.2) with $k \in I$, we have $I^t = I = p^{-1}(\mathbb{S})$ and $\operatorname{sign}(\langle \psi_k, y\rangle) = \sigma(k)$.

To ensure that $I^t = I$, this means the recovery of all $i \in I$, we need to have

$$\min_{i \in I} |\langle \psi_i, y\rangle| > \max_{i \notin I} |\langle \psi_i, y\rangle|. \tag{3.33}$$

Expanding the inner product of a rescaled signal y with an atom ψ_i of the perturbed dictionary Ψ yields

$$\begin{aligned} |\langle \psi_i, \Phi x_{c,p,\sigma} + r\rangle| &= |\sum_j \sigma(j) c(p(j)) \langle \psi_i, \phi_j\rangle + \langle \psi_i, r\rangle| \\ &= |c(p(i))\langle \psi_i, \phi_i\rangle + \sigma(i) \sum_{j \neq i} \sigma(j) c(p(j)) \langle \psi_i, \phi_j\rangle + \sigma(i)\langle \psi_i, r\rangle|. \end{aligned}$$

Depending on the index i under consideration, we obtain the following bounds from below resp. above,

$$\begin{aligned} i \in I: &\quad |\langle \psi_i, \Phi x_{c,p,\sigma} + r\rangle| \geq c(S)\alpha_{\min} - \big|\sum_{j \neq i} \sigma(j) c(p(j)) \langle \psi_i, \phi_j\rangle\big| - |\langle \psi_i, r\rangle|, \\ i \notin I: &\quad |\langle \psi_i, \Phi x_{c,p,\sigma} + r\rangle| \leq c(S+1)\alpha_{\max} + \big|\sum_{j \neq i} \sigma(j) c(p(j)) \langle \psi_i, \phi_j\rangle\big| + |\langle \psi_i, r\rangle|. \end{aligned}$$

This means that a sufficient condition for the recovery of I is that for all i

$$\big|\sum_{j\neq i}\sigma(j)c(p(j))\langle\psi_i,\phi_j\rangle\big| < \theta_1\cdot c(S)\alpha_{\min} \quad \text{and} \quad |\langle\psi_i,r\rangle| < \theta_2\cdot c(S)\alpha_{\min}, \tag{3.34}$$

where θ_1 and θ_2 ensure that

$$\begin{aligned} c(S)\alpha_{\min} - &\theta_1 c(S)\alpha_{\min} - \theta_2 c(S)\alpha_{\min} \\ &\overset{!}{\geq} c(S{+}1)\alpha_{\max} + \theta_1 c(S)\alpha_{\min} + \theta_2 c(S)\alpha_{\min}. \end{aligned} \tag{3.35}$$

Since the conditions above also guarantee the recovery of the correct sign $\sigma(i)$ for all $i\in I$, so in particular the recovery of $\sigma(k)$, we can bound the probability of the event that thresholding fails while k is in the generating support I as

$$\begin{aligned} &\mathbb{P}\left(\left[I^t\neq I\vee\operatorname{sign}(\langle\psi_k,y\rangle)\neq\sigma(k)\right]\wedge k\in I\right) \\ &\qquad\leq \mathbb{P}\big(\exists i: |\sum_{j\neq i}\sigma(j)c(p(j))\langle\psi_i,\phi_j\rangle| \geq \theta_1\cdot c(S)\alpha_{\min}\wedge k\in I\big) \\ &\qquad\qquad + \mathbb{P}\big(\exists i: |\langle\psi_i,r\rangle| \geq \theta_2\cdot c(S)\alpha_{\min}\wedge k\in I\big) \\ &\qquad\leq \sum_i \mathbb{P}\Big(\big|\sum_{j\neq i}\sigma(j)c(p(j))\langle\psi_i,\phi_j\rangle\big| \geq \theta_1 c(S)\alpha_{\min}\wedge k\in I\Big) \\ &\qquad\qquad + \sum_i \mathbb{P}\big(|\langle\psi_i,r\rangle| \geq \theta_2 c(S)\alpha_{\min}\wedge k\in I\big) \\ &\qquad\leq \sum_i\sum_{\ell\in\mathbb{S}} \mathbb{P}\Big(\big|\sum_{j\neq i}\sigma(j)c(p(j))\langle\psi_i,\phi_j\rangle\big| \geq \theta_1 c(S)\alpha_{\min}\big|p(k)=\ell\Big)\cdot\mathbb{P}(p(k)=\ell) \\ &\qquad\qquad + \sum_i\sum_{\ell\in\mathbb{S}} \mathbb{P}\big(|\langle\psi_i,r\rangle| \geq \theta_2 c(S)\alpha_{\min}\big|p(k)=\ell\big)\cdot\mathbb{P}(p(k)=\ell). \end{aligned}$$

Since every permutation is equally likely, each index is equally likely to be mapped to ℓ, meaning $\mathbb{P}(p(k)=\ell)=1/K$. Using the independence of the noise from the remaining signal parameters and its sub-Gaussian property further leads to

$$\begin{aligned} &\mathbb{P}\left(\left[I^t\neq I\vee\operatorname{sign}(\langle\psi_k,y\rangle)\neq\sigma(k)\right]\wedge k\in I\right) \\ &\qquad\leq \frac{1}{K}\sum_i\sum_{\ell\in\mathbb{S}} \mathbb{P}\Big(\big|\sum_{j\neq i}\sigma(j)c(p(j))\langle\psi_i,\phi_j\rangle\big| \geq \theta_1 c(S)\alpha_{\min}\big|p(k)=\ell\Big) \\ &\qquad\qquad + \frac{S}{K}\sum_i \mathbb{P}\big(|\langle\psi_i,r\rangle| \geq \theta_2 c(S)\alpha_{\min}\big) \\ &\qquad\leq \frac{1}{K}\sum_i\sum_{\ell\in\mathbb{S}} \mathbb{P}\Big(\big|\sum_{j\neq i}\sigma(j)c(p(j))\langle\psi_i,\phi_j\rangle\big| \geq \theta_1 c(S)\alpha_{\min}\big|p(k)=\ell\Big) \\ &\qquad\qquad + 2S\exp\left(\frac{-(\theta_2 c(S)\alpha_{\min})^2}{2\rho^2}\right). \end{aligned} \tag{3.36}$$

To estimate the terms $\mathbb{P}\big(\big|\sum_{j\neq i}\sigma(j)c(p(j))\langle\psi_i,\phi_j\rangle\big|\geq\theta_1 c(S)\alpha_{\min}\big|p(k)=\ell\big)$, we split the sum into two parts, one over $j\in I\backslash\{i,k\}$, that captures most of the energy, and the other over $j\in(I^c\cup\{k\})\backslash\{i\}$. For $m_1\in(0,1)$ and $m_2=1-m_1$, we have

$$\begin{aligned}
&\mathbb{P}\Big(\big|\sum_{j\neq i}\sigma(j)c(p(j))\langle\psi_i,\phi_j\rangle\big|\geq\theta_1 c(S)\alpha_{\min}\big|p(k)=\ell\Big)\\
&\quad\leq\mathbb{P}\Big(\big|\sum_{j\in I\backslash\{i,k\}}\sigma(j)c(p(j))\langle\psi_i,\phi_j\rangle\big|\\
&\qquad+\big|\sigma(k)c(p(k))\langle\psi_i,\phi_k\rangle+\sum_{j\in I^c\backslash\{i\}}\sigma(j)c(p(j))\langle\psi_i,\phi_j\rangle\big|\geq\theta_1\cdot c(S)\alpha_{\min}\big|p(k)=l\Big)\\
&\quad\leq\mathbb{P}\Big(\big|\sum_{j\in I\backslash\{i,k\}}\sigma(j)c(p(j))\langle\psi_i,\phi_j\rangle\big|\geq m_1\theta_1 c(S)\alpha_{\min}\big|p(k)=\ell\Big)\\
&\qquad\quad+\mathbb{P}\Big(\big|\sum_{j\in(I^c\cup\{k\})\backslash\{i\}}\sigma(j)c(p(j))\langle\psi_i,\phi_j\rangle\big|\geq m_2\theta_1 c(S)\alpha_{\min}\big|p(k)=\ell\Big).
\end{aligned}$$

The first term on the right hand side we estimate using Proposition 3.4 and the second term using Hoeffding's inequality. Depending on the the index i under consideration, we get the following bounds. If $i=k$ we have

$$\begin{aligned}
&\mathbb{P}\Big(\big|\sum_{j\neq k}\sigma(j)c(p(j))\langle\psi_k,\phi_j\rangle\big|\geq\theta_1 c(S)\alpha_{\min}\big|p(k)=\ell\Big)\\
&\quad\leq\mathbb{P}\Big(\big|\sum_{j\in I\backslash\{k\}}\sigma(j)c(p(j))\langle\psi_k,\phi_j\rangle\big|\geq m_1\theta_1 c(S)\alpha_{\min}\big|p(k)=\ell\Big)\\
&\qquad\quad+\mathbb{P}\Big(\big|\sum_{j\in I^c}\sigma(j)c(p(j))\langle\psi_k,\phi_j\rangle\big|\geq m_2\theta_1 c(S)\alpha_{\min}\big|p(k)=\ell\Big)\\
&\quad\leq 2\exp\left(\frac{-(m_1\theta_1 c(S)\alpha_{\min})^2}{2(c(1)\hat{\mu}\cdot m_1\theta_1 c(S)\alpha_{\min}+\|c(\mathbb{S})\|_2^2\frac{B}{K-S})}\right)+2\exp\left(\frac{-(m_2\theta_1 c(S)\alpha_{\min})^2}{2\hat{\mu}^2\|c(\mathbb{S}^c)\|_2^2}\right),
\end{aligned}$$

where we used that $\sum_{j\in I^c}c(p(j))^2\left|\langle\psi_k,\phi_j\rangle\right|^2\leq\hat{\mu}^2\|c(\mathbb{S}^c)\|_2^2$. Note that the sign sequence σ is independent of the permutation p and hence, we can apply Hoeffding's inequality also to the conditional probability. More precisely, the expectation used in the above inequality is only over σ, independent from p. Note also, the residual energy $\|c(\mathbb{S}^c)\|_2^2$ is zero for perfectly S-sparse signals and can be assumed small otherwise. In case of $\|c(\mathbb{S}^c)\|_2^2=0$, the sum over I^c is zero and hence, the last term vanishes.

Similarly, if $i \neq k$ we get

$$\begin{aligned}\mathbb{P}\Big(\big|\sum_{j\neq i}\sigma(j)c(p(j))\langle\psi_i,\phi_j\rangle\big| \geq \theta_1 c(S)\alpha_{\min}\big|p(k)=\ell\Big)&\\ \leq 2\exp\left(\frac{-(m_1\theta_1 c(S)\alpha_{\min})^2}{2(c(1)\hat{\mu}\cdot m_1\theta_1 c(S)\alpha_{\min}+\|c(\mathbb{S})\|_2^2\frac{B}{K-S})}\right)&\\ +2\exp\left(\frac{-(m_2\theta_1 c(S)\alpha_{\min})^2}{2\hat{\mu}^2(c(\ell)^2+\|c(\mathbb{S}^c)\|_2^2)}\right)&.\end{aligned}$$

Hence, with some small simplifications we get for all i, including k,

$$\begin{aligned}&\mathbb{P}\Big(\big|\sum_{j\neq i}\sigma(j)c(p(j))\langle\psi_i,\phi_j\rangle\big| \geq \theta_1 c(S)\alpha_{\min}\big|p(k)=\ell\Big)\\ &\leq 2\exp\left(\frac{-(m_1\theta_1 c(S)\alpha_{\min})^2}{2(c(1)\hat{\mu}\cdot m_1\theta_1 c(S)\alpha_{\min}+\|c(\mathbb{S})\|_2^2\frac{B}{K-S})}\right)\\ &\qquad +2\exp\left(\frac{-(m_2\theta_1 c(S)\alpha_{\min})^2}{2\hat{\mu}^2(c(\ell)^2+\|c(\mathbb{S}^c)\|_2^2)}\right)\\ &\leq 2\exp\left(-\frac{1}{4}\min\left\{\frac{c(S)m_1\theta_1\alpha_{\min}}{c(1)\hat{\mu}},\frac{(K-S)(m_1\theta_1 c(S)\alpha_{\min})^2}{B\|c(\mathbb{S})\|_2^2}\right\}\right)\\ &\qquad +2\exp\left(-\frac{1}{4}\min\left\{\frac{(c(S)m_2\theta_1\alpha_{\min})^2}{c(1)^2\hat{\mu}^2},\frac{(c(S)m_2\theta_1\alpha_{\min})^2}{\hat{\mu}^2\|c(\mathbb{S}^c)\|_2^2}\right\}\right).\end{aligned}$$

Substituting the expression above into (3.36), we get

$$\begin{aligned}&\mathbb{P}\left(\left[I^t\neq I\vee\operatorname{sign}(\langle\psi_k,y\rangle)\neq\sigma(k)\right]\wedge k\in I\right)\\ &\leq 2S\exp\left(-\frac{1}{4}\min\left\{\frac{c(S)m_1\theta_1\alpha_{\min}}{c(1)\hat{\mu}},\frac{(K-S)(m_1\theta_1 c(S)\alpha_{\min})^2}{B\|c(\mathbb{S})\|_2^2}\right\}\right)\\ &\qquad +2S\exp\left(-\frac{1}{4}\min\left\{\frac{(c(S)m_2\theta_1\alpha_{\min})^2}{c(1)^2\hat{\mu}^2},\frac{(c(S)m_2\theta_1\alpha_{\min})^2}{\hat{\mu}^2\|c(\mathbb{S}^c)\|_2^2}\right\}\right)\\ &\qquad +2S\exp\left(\frac{-(\theta_2 c(S)\alpha_{\min})^2}{2\rho^2}\right),\end{aligned}$$

where θ_1 and θ_2 have to ensure (3.35) and $m_1+m_2=1$. From this, whenever

$$\begin{aligned}\alpha_{\min}\geq\max\Bigg\{&\frac{1}{1-2\theta_1-2\theta_2}\frac{c(S+1)}{c(S)}\alpha_{\max},\ \frac{4n}{m_1\theta_1}\frac{c(1)}{c(S)}\hat{\mu}\log K,\\ &\frac{2\sqrt{n}}{m_1\theta_1}\frac{\|c(\mathbb{S})\|_2}{c(S)}\sqrt{\frac{B\log K}{K-S}},\ \frac{2\sqrt{n}}{(1-m_1)\theta_1}\frac{c(1)}{c(S)}\hat{\mu}\sqrt{\log K},\\ &\frac{2\sqrt{n}}{(1-m_1)\theta_1}\frac{\|c(\mathbb{S}^c)\|_2}{c(S)}\hat{\mu}\sqrt{\log K},\ \frac{\sqrt{2n}}{\theta_2}\frac{\rho}{c(S)}\sqrt{\log K}\Bigg\},\end{aligned}$$

we get that

$$\mathbb{P}\left(\left[I^t \neq I \vee \operatorname{sign}(\langle\psi_k, y\rangle) \neq \sigma(k)\right] \wedge k \in I\right) \leq 6S \cdot K^{-n}.$$

Setting $\theta_1 = \frac{6}{16}$, $\theta_2 = \frac{1}{16}$, $m_1 = \frac{2}{3}$, $n = 3$, the probability that thresholding fails to recover I and/or the corresponding signs, restricted to the signals for which we have $k \in I$, is bounded by $6S \cdot K^{-3}$, whenever

$$\alpha_{\min} \geq \max\left\{8\frac{c(S{+}1)}{c(S)}\alpha_{\max},\ 48\frac{c(1)}{c(S)}\hat{\mu}\log K,\ 14\frac{c(1)}{c(S)}\sqrt{\frac{SB\log K}{K-S}},\ 40\frac{\rho}{c(S)}\sqrt{\log K}\right\},$$

and $\frac{\|c(\mathbb{S}^c)\|_2}{c(1)} \leq \frac{12}{7}\sqrt{\log K}$, where we have used that $\|c(\mathbb{S})\|_2 \leq \sqrt{S}c(1)$.

Step 2 - Probability of wrongly recovering k for $k \notin I$ - $\mathbb{P}(k \in I^t | k \notin I)$

As a second step we bound the probability of wrongly recovering an atom ψ_k when it is not in the generating support, meaning $k \notin I$. A sufficient condition for not recovering k is that

$$\min_{i\in I} |\langle\psi_i, y\rangle| > |\langle\psi_k, y\rangle|. \tag{3.37}$$

Using the bounds from step 1,

$$\begin{aligned} i \in I: \quad & |\langle\psi_i, \Phi x_{c,p,\sigma} + r\rangle| \geq c(S)\alpha_{\min} - \Big|\sum_{j\neq i}\sigma(j)c(p(j))\langle\psi_i,\phi_j\rangle\Big| - |\langle\psi_i, r\rangle|, \\ k \notin I: \quad & |\langle\psi_k, \Phi x_{c,p,\sigma} + r\rangle| \leq c(S{+}1)\alpha_k + \Big|\sum_{j\neq k}\sigma(j)c(p(j))\langle\psi_k,\phi_j\rangle\Big| + |\langle\psi_k, r\rangle|, \end{aligned}$$

we get as sufficient condition for not recovering k, that for all $i \in I \cup \{k\}$

$$\Big|\sum_{j\neq i}\sigma(j)c(p(j))\langle\psi_i,\phi_j\rangle\Big| < \theta_1 \cdot c(S)\alpha_{\min} \quad \text{and} \quad |\langle\psi_i, r\rangle| < \theta_2 \cdot c(S)\alpha_{\min},$$

where θ_1 and θ_2 again ensure that

$$c(S)\alpha_{\min} - \theta_1 c(S)\alpha_{\min} - \theta_2 c(S)\alpha_{\min} \overset{!}{\geq} c(S{+}1)\alpha_k + \theta_1 c(S)\alpha_{\min} + \theta_2 c(S)\alpha_{\min}.$$

We now bound the probability of thresholding recovering k when it is not in the generating support I as

$$\begin{aligned}
&\mathbb{P}(k \in I^t \wedge k \notin I)\\
&\qquad = \sum_{\ell > S} \mathbb{P}(k \in I^t \big| p(k) = \ell) \cdot \mathbb{P}(p(k) = \ell)\\
&\qquad \le \frac{1}{K} \sum_{\ell > S} \sum_{i \in I \cup \{k\}} \mathbb{P}\Big(\big| \sum_{j \neq i} \sigma(j) c(p(j)) \langle \psi_i, \phi_j \rangle \big| \ge \theta_1 \cdot c(S) \alpha_{\min} \big| p(k) = \ell \Big)\\
&\qquad\qquad + \frac{1}{K} \sum_{\ell > S} \sum_{i \in I \cup \{k\}} \mathbb{P}\big(|\langle \psi_i, r \rangle| \ge \theta_2 \cdot c(S) \alpha_{\min} \big| p(k) = \ell \big).
\end{aligned}$$

Using the same splitting technique as in step 1 and the sub-Gaussian property of r, we get

$$\begin{aligned}
&\mathbb{P}(k \in I^t \wedge k \notin I)\\
&\quad \le 2(S+1) \exp\left(-\frac{1}{4} \min\left\{ \frac{c(S) m_1 \theta_1 \alpha_{\min}}{c(1)\hat{\mu}}, \frac{(K-S)(m_1 \theta_1 c(S) \alpha_{\min})^2}{B \| c(\mathbb{S}) \|_2^2} \right\} \right)\\
&\qquad + 2(S+1) \exp\left(\frac{-(m_2 \theta_1 c(S) \alpha_{\min})^2}{2 \hat{\mu}^2 \| c(\mathbb{S}^c) \|_2^2} \right) + 2(S+1) \exp\left(\frac{-(\theta_2 c(S) \alpha_{\min})^2}{2\rho^2} \right).
\end{aligned}$$

In order to have this probability sufficiently small, we need to have

$$\begin{aligned}
\alpha_{\min} \ge \max \Bigg\{ &\frac{1}{1 - 2\theta_1 - 2\theta_2} \frac{c(S{+}1)}{c(S)} \alpha_k,\ \frac{4n}{m_1 \theta_1} \frac{c(1)}{c(S)} \hat{\mu} \log K,\ \frac{2\sqrt{n}}{m_1 \theta_1} \frac{\| c(\mathbb{S}) \|_2}{c(S)} \sqrt{\frac{B \log K}{K - S}},\\
&\frac{\sqrt{2n}}{(1 - m_1)\theta_1} \frac{\| c(\mathbb{S}^c) \|_2}{c(S)} \hat{\mu} \sqrt{\log K},\ \frac{\sqrt{2n}}{\theta_2} \frac{\rho}{c(S)} \sqrt{\log K} \Bigg\}.
\end{aligned}$$

Choosing the same values as before, $\theta_1 = \frac{6}{16}$, $\theta_2 = \frac{1}{16}$, $m_1 = \frac{2}{3}$, $n = 3$, we arrive at the bound

$$\mathbb{P}(k \in I^t \wedge k \notin I) \le 6(S+1) \cdot K^{-3},$$

whenever $\frac{\| c(\mathbb{S}^c) \|_2}{c(1)} \le \frac{12}{5} \sqrt{\log K}$ and

$$\alpha_{\min} \ge \max \left\{ 8 \frac{c(S{+}1)}{c(S)} \alpha_k,\ 48 \frac{c(1)}{c(S)} \hat{\mu} \log K,\ 14 \frac{c(1)}{c(S)} \sqrt{\frac{SB \log K}{K - S}},\ 40 \frac{\rho}{c(S)} \sqrt{\log K} \right\}.$$

Using all these estimates, we are now ready to bound the error originating from the difference between the thresholding and the oracle residual.

Step 3 - Putting it all together

Using all previous estimates, what remains to do is to estimate the size of $\mathcal{F}$ and $\mathcal{G}$ and finally put all pieces together. Inserting our probability estimates into (3.31) and (3.32), we get

$$\mathbb{P}\left(|\mathcal{F}| \geq N\left(\frac{6S}{K^3} + \frac{C_r\gamma_{1,S}}{3K\sqrt{B+1}}t\varepsilon\right)\right) \leq \exp\left(-\frac{NC_r^2\gamma_{1,S}^2t^2\varepsilon^2}{\frac{108S(B+1)}{K} + 3t\varepsilon C_r\gamma_{1,S}K\sqrt{B+1}}\right)$$

and

$$\mathbb{P}\left(|\mathcal{G}| \geq N\left(\frac{6(S+1)}{K^3} + \frac{C_r\gamma_{1,S}}{3K\sqrt{B+1}}t\varepsilon\right)\right) \leq \exp\left(-\frac{NC_r^2\gamma_{1,S}^2t^2\varepsilon^2}{\frac{108(S+1)(B+1)}{K} + 3t\varepsilon C_r\gamma_{1,S}K\sqrt{B+1}}\right),$$

respectively. As we have

$$\Big\| \sum_n \left[R^t(\Psi, y_n, k) - R^o(\Psi, y_n, k)\right] \Big\|_2 \leq (2|\mathcal{F}| + |\mathcal{G}|)\sqrt{B+1},$$

in summary, we get

$$\mathbb{P}\left(\frac{1}{N}\Big\| \sum_n \left[R^t(\Psi, y_n, k) - R^o(\Psi, y_n, k)\right] \Big\|_2 > \frac{18(S+1)\sqrt{B+1}}{K^3} + \frac{C_r\gamma_{1,S}}{K}t\varepsilon\right) \leq 2\exp\left(-\frac{NC_r^2\gamma_{1,S}^2t^2\varepsilon^2}{\frac{108(S+1)(B+1)}{K} + 3t\varepsilon C_r\gamma_{1,S}K\sqrt{B+1}}\right).$$

□

Next we will prove the lemma yielding a bound for the error originating from the difference between the oracle residuals based on the generating dictionary and a perturbation of it. For our estimates we will use some results of [61], adapted to our problem and with some slight modifications.

3.2.2 Difference between oracle residuals

For the proof of Lemma 3.7 we will use the vector version of Bernstein's inequality.

Theorem 3.6 (Vector Bernstein, [34, 25, 35]). *Let $(v_n)_n \in \mathbb{R}^d$ be a finite sequence of independent random vectors. If $\|v_n\|_2 \leq M$ almost surely, $\|\mathbb{E}(v_n)\|_2 \leq m_1$ and $\sum_n \mathbb{E}(\|v_n\|_2^2) \leq m_2$, then for all $0 \leq t \leq m_2/(M+m_1)$, we have*

$$\mathbb{P}\left(\Big\|\sum_n v_n - \sum_n \mathbb{E}(v_n)\Big\|_2 \geq t\right) \leq \exp\left(-\frac{t^2}{8m_2} + \frac{1}{4}\right), \tag{3.38}$$

and, in general,

$$\mathbb{P}\left(\Big\|\sum_n v_n - \sum_n \mathbb{E}(v_n)\Big\|_2 \geq t\right) \leq \exp\left(-\frac{t}{8}\cdot \min\left\{\frac{t}{m_2}, \frac{1}{M+m_1}\right\} + \frac{1}{4}\right). \qquad (3.39)$$

Note that the general statement is simply a consequence of the first part, since for $t \geq m_2/(M+m_1)$ we can choose $m_2 = t(M+m_1)$.

In the following we prove that, assuming incoherence and good conditioning of the perturbed dictionary, the oracle residuals based on the perturbed dictionary Ψ and the generating dictionary Φ are close to each other.

Lemma 3.7. *Assume that the signals y_n follow the random model in* (3.2)*. Further, assume that $S \leq \min\left\{\frac{K}{98\|\Phi\|_{2,2}^2}, \frac{1}{98\rho^2}\right\}$ and that the current estimate of the dictionary Ψ has distance $d(\Phi,\Psi) = \varepsilon \geq \frac{1}{32\sqrt{S}}$ but is incoherent and well conditioned, meaning its coherence $\mu(\Psi)$ and its operator norm $\|\Psi\|_{2,2}$ satisfy*

$$\mu(\Psi) \leq \frac{1}{20\log K} \quad \text{and} \quad \|\Psi\|_{2,2}^2 \leq \frac{K}{134e^2 S\log K} - 1. \qquad (3.40)$$

Then for all $0 \leq t \leq 1/8$ we have

$$\mathbb{P}\left(\frac{1}{N}\Big\|\sum_n [R^o(\Psi, y_n, k) - R^o(\Phi, y_n, k)]\Big\|_2 \geq \frac{C_r\gamma_{1,S}}{K}(0.308\varepsilon + t\varepsilon)\right)$$
$$\leq \exp\left(-\frac{NC_r^2\gamma_{1,S}^2 t^2\varepsilon}{12K\max\{S, \|\Phi\|_{2,2}^2\}^{\frac{3}{2}}} + \frac{1}{4}\right).$$

Proof. Throughout the proof we use the abbreviations $B = \|\Phi\|_{2,2}^2$, $\bar{B} = \|\Psi\|_{2,2}^2$, $\mu = \mu(\Phi)$ and $\bar{\mu} = \mu(\Psi)$. As in [61], we apply Theorem 3.6 to $v_n = R^o(\Psi, y_n, k) - R^o(\Phi, y_n, k)$, and drop the index n for conciseness. From Lemma B.8 in [61] we have that $v = T(I,k)y \cdot \sigma(k)\cdot \chi(I,k)$, where $T(I,k) := P(\Phi_I) - P(\Psi_I) - P(\phi_k) + P(\psi_k)$, and for its expectation,

$$\mathbb{E}(v) = \frac{C_r\gamma_{1,S}}{K}\binom{K-1}{S-1}^{-1}\sum_{|I|=S, k\in I}\big[P(\psi_k) - P(\Psi_I)\big]\phi_k. \qquad (3.41)$$

Using the orthogonal decomposition $\phi_k = [P(\psi_k) + Q(\psi_k)]\phi_k$, where $P(\psi_k)Q(\psi_k) = 0$, we get

$$\mathbb{E}(v) = \frac{C_r\gamma_{1,S}}{K}\binom{K-1}{S-1}^{-1}\sum_{|I|=S, k\in I} -P(\Psi_I)Q(\psi_k)\phi_k. \qquad (3.42)$$

Since the perturbed dictionary Ψ is well-conditioned and incoherent, for most I the subdictionary Ψ_I will be a quasi isometry and $\|\Psi_I\Psi_I^\star - P(\Psi_I)\|_{2,2} \leq \delta(\Psi_I) \leq \delta_0$. We therefore expand the expectation above, using the abbreviation $p_{K,S} = \binom{K-1}{S-1}^{-1}$, as

$$\begin{aligned}
\frac{K}{C_r\gamma_{1,S}}\,\mathbb{E}(v) &= p_{K,S}\left(\sum_{|I|=S,k\in I}\big[\Psi_I\Psi_I^\star - P(\Psi_I)\big]Q(\psi_k)\phi_k - \sum_{|I|=S,k\in I}\Psi_{I\setminus k}\Psi_{I\setminus k}^\star Q(\psi_k)\phi_k\right)\\
&= p_{K,S}\left(\sum_{|I|=S,k\in I}\big[\Psi_I\Psi_I^\star - P(\Psi_I)\big]Q(\psi_k)\phi_k - \binom{K-2}{S-2}\sum_{j\neq k}\psi_j\psi_j^\star Q(\psi_k)\phi_k\right)\\
&= p_{K,S}\sum_{|I|=S,k\in I}\big[\Psi_I\Psi_I^\star - P(\Psi_I)\big]Q(\psi_k)\phi_k - \frac{S-1}{K-1}(\Psi\Psi^\star - \psi_k\psi_k^\star)Q(\psi_k)\phi_k\\
&= p_{K,S}\sum_{\substack{|I|=S,k\in I\\ \delta(\Psi_I)\leq\delta_0}}\big[\Psi_I\Psi_I^\star - P(\Psi_I)\big]Q(\psi_k)\phi_k\\
&\qquad + p_{K,S}\sum_{\substack{|I|=S,k\in I\\ \delta(\Psi_I)>\delta_0}}\big[\Psi_I\Psi_I^\star - P(\Psi_I)\big]Q(\psi_k)\phi_k - \frac{S-1}{K-1}\Psi\Psi^\star Q(\psi_k)\phi_k.
\end{aligned}$$

Since for $\psi_k = \alpha_k\phi_k + \omega_k z_k$, we have $\|Q(\psi_k)\phi_k\|_2 = \omega_k \leq \varepsilon$, we can bound the norm of the expectation above as

$$\|\mathbb{E}(v)\|_2 \leq \frac{C_r\gamma_{1,S}}{K}\left[\delta_0 + \mathbb{P}\big(\delta(\Psi_I) > \delta_0\big||I| = S, k\in I\big)\cdot(\bar{B}+1) + \frac{(S-1)\bar{B}}{K-1}\right]\varepsilon. \quad (3.43)$$

In order to estimate the probability of a subdictionary being ill-conditioned we use Chretien and Darses's result on the conditioning of random subdictionaries, which is slightly cleaner and thus easier to handle than the original result by Tropp, [66]. Hence, using Theorem 3.1 of [14], reformulated for our purposes, we get

$$\begin{aligned}
\mathbb{P}\left(\delta(\Psi_I) > \delta_0 \,\big||I| = S\right) &\leq 216K\cdot\exp\left(-\min\left\{\frac{\delta_0}{2\bar{\mu}}, \frac{\delta_0^2 K}{4e^2 S\bar{B}}\right\}\right)\\
&\leq 216K\cdot\max\{K^{-n_1}, K^{-n_2}\},
\end{aligned}$$

whenever,

$$\bar{\mu} \leq \frac{\delta_0}{2n_1\log K} \quad\text{and}\quad \bar{B} \leq \frac{\delta_0^2 K}{4n_2 e^2 S\log K}.$$

Together with the union bound,

$$\begin{aligned}\mathbb{P}(\delta(\Psi_I) > \delta_0 \,\big|\, |I| = S, k \in I) &= \binom{K-1}{S-1}^{-1} |\{I : \delta(\Psi_I) > \delta_0, |I| = S, k \in I\}| \\ &\leq \binom{K-1}{S-1}^{-1} |\{I : \delta(\Psi_I) > \delta_0, |I| = S\}| \\ &= \frac{K}{S} \cdot \mathbb{P}\left(\delta(\Psi_I) > \delta_0 \,\big|\, |I| = S\right),\end{aligned}$$

this leads to

$$\mathbb{P}(\delta(\Psi_I) > \delta_0 \,\big|\, |I| = S, k \in I) \leq 216\frac{K^2}{S} \cdot \max\{K^{-n_1}, K^{-n_2}\}.$$

Substituting this bound into (3.43), we obtain

$$\|\mathbb{E}(v)\|_2 \leq \frac{C_r \gamma_{1,S}}{K}\left[\delta_0 + 216\frac{K^2(\bar{B}+1)}{S} \cdot \max\{K^{-n_1}, K^{-n_2}\} + \frac{S\bar{B}}{K}\right]\varepsilon.$$

Choosing $\delta_0 = 3/10$, $n_1 = n_2 = 3$, as long as $\bar{B} \leq \frac{K}{134e^2 S \log K} - 1$ and $\mu(\Psi) \leq \frac{1}{20 \log K}$, we have

$$\begin{aligned}\|\mathbb{E}(v)\|_2 &\leq \frac{C_r \gamma_{1,S}}{K}\left[\frac{3}{10} + \frac{216}{134e^2 S^2 \log K} + \frac{1}{134e^2 \log K}\right]\varepsilon \\ &\leq 0.308 \cdot \frac{C_r \gamma_{1,S}}{K}\varepsilon, \qquad (3.44)\end{aligned}$$

where we used that $S \geq 2$ and $\log K \geq 7$.

The second quantity we need to bound is the expected energy of $v = T(I,k)y \cdot \sigma(k) \cdot \chi(I,k)$. Combining Eqs. (115-118) from Lemma B.8 in [61], we get

$$\mathbb{E}(\|v\|_2^2) \leq \mathbb{E}_p\left(\chi(I,k)\left[4\gamma_{2,S}\varepsilon^2 + \left(\frac{B(1-\gamma_{2,S})}{K-S} + \rho^2\right)\|T(I,k)\|_F^2\right]\right). \qquad (3.45)$$

From this we see, what remains to do is to bound the norm term $\|T(I,k)\|_F^2$. As we only consider the case where we have $\varepsilon \geq \frac{1}{32\sqrt{S}}$, we will use a crude but painless estimate and in turn accept an additional factor S in the final sample complexity. Concretely, as $T(I,k)$ is the difference of two orthogonal projections onto subspaces of dimension $S-1$, namely $P(\Phi_I) - P(\phi_k)$ and $P(\Psi_I) - P(\psi_k)$, we get

$$\|T(I,k)\|_F^2 = \|P(\Phi_I) - P(\phi_k) - [P(\Psi_I) - P(\psi_k)]\|_F^2 \leq 2(S-1),$$

and therefore,

$$\begin{aligned}\mathbb{E}\left(\|v\|_2^2\right) &\leq \frac{S}{K}\left(4\gamma_{2,S}\varepsilon^2 + 2(S-1)\left(\frac{B(1-\gamma_{2,S})}{K-S} + \rho^2\right)\right)\\ &\leq \frac{S}{K}\left(4\gamma_{2,S}\varepsilon^2 + \frac{2SB(1-\gamma_{2,S})}{K-S} + 2S\rho^2\right)\\ &\leq \frac{S}{K}\left(4\varepsilon^2 + \frac{1}{24}\right),\end{aligned}$$

where for the last inequality we have used the assumption $S \leq \min\{\frac{K}{98B}, \frac{1}{98\rho^2}\}$. Combining the estimates for $\|\mathbb{E}(v)\|_2$ and $\mathbb{E}(\|v\|_2^2)$ with the norm bound

$$\|v\|_2 = \|\,[P(\Phi_I) - P(\Psi_I) - P(\phi_k) + P(\psi_k)]\,y\|_2 \leq 2\|y\|_2 \leq 2\sqrt{B+1},$$

we get for the case where $\varepsilon \geq \frac{1}{32\sqrt{S}}$ and $0 \leq t \leq 1/8$,

$$\begin{aligned}&\mathbb{P}\left(\Big\|\sum_n [v_n - \mathbb{E}(v_n)]\Big\|_2 \geq \frac{C_r\gamma_{1,S}}{K}t\varepsilon\right)\\ &\qquad\leq \exp\left(-\frac{C_r\gamma_{1,S}t\varepsilon}{8K}\cdot\min\left\{\frac{C_r\gamma_{1,S}t\varepsilon}{S(4\varepsilon^2+1/24)}, \frac{1}{\varepsilon+2\sqrt{B+1}}\right\} + \frac{1}{4}\right)\\ &\qquad\leq \exp\left(-\frac{C_r^2\gamma_{1,S}^2t^2\varepsilon}{8K}\cdot\min\left\{\frac{1}{S(4\varepsilon+(24\varepsilon)^{-1})}, \frac{1}{3t\gamma_{1,S}\sqrt{B+1}}\right\} + \frac{1}{4}\right)\\ &\qquad\leq \exp\left(-\frac{C_r^2\gamma_{1,S}^2t^2\varepsilon}{8K\max\{S,B\}}\cdot\min\left\{\frac{1}{4\varepsilon+(24\varepsilon)^{-1}}, \frac{1}{3t\sqrt{2}}\right\} + \frac{1}{4}\right)\\ &\qquad\leq \exp\left(-\frac{C_r^2\gamma_{1,S}^2t^2\varepsilon}{12K\max\{S,B\}^{\frac{3}{2}}} + \frac{1}{4}\right),\end{aligned}$$

where we have used that $C_r \leq 1$, $\gamma_{1,S} \leq \sqrt{S}$, $\varepsilon \leq \sqrt{2}$ and $B+1 \geq 2$. Using that we have

$$\begin{aligned}\frac{1}{N}\Big\|\sum_n [R^o(\Psi, y_n, k) - R^o(\Phi, y_n, k)]\Big\|_2 &= \frac{1}{N}\Big\|\sum_n [v_n - \mathbb{E}(v_n) + \mathbb{E}(v_n)]\Big\|_2\\ &\leq \frac{1}{N}\Big\|\sum_n [v_n - \mathbb{E}(v_n)]\Big\|_2 + \|\mathbb{E}(v_n)\|_2,\end{aligned}$$

we finally get for $\varepsilon \geq \frac{1}{32\sqrt{S}}$ and $0 \leq t \leq 1/8$,

$$\begin{aligned}&\mathbb{P}\left(\frac{1}{N}\Big\|\sum_n [R^o(\Psi, y_n, k) - R^o(\Phi, y_n, k)]\Big\|_2 \geq \frac{C_r\gamma_{1,S}}{K}(0.308\varepsilon + t\varepsilon)\right)\\ &\qquad\qquad\leq \exp\left(-\frac{NC_r^2\gamma_{1,S}^2t^2\varepsilon}{12K\max\{S,B\}^{\frac{3}{2}}} + \frac{1}{4}\right).\end{aligned}$$

□

In the next chapter we will analyse situations in which ITKrM does not recover the generating dictionary. For that, we will first run some numerical experiments showing that in such situations ITKrM produces a dictionary without the cross-coherence property from Theorem 3.1. Further, we will also analyse why it is hard to escape from such dictionaries. This will finally lead us to some interesting observations that open the road to further improve the convergence behaviour of ITKrM and the automatic choice of the sparsity level S and the dictionary size K.

Chapter 4

Beyond the Contractive Areas - Replacement and Adaptivity

In this chapter, we analyse the behaviour of ITKrM outside the areas where it is a contraction. This will show us that there seem to exist stable fixed points which are not equivalent to the generating dictionary and can be characterised as very coherent. Based on a closer inspection of the residuals at these spurious fixed points we develop a replacement strategy and a strategy to find good replacement candidates that allow ITKrM to escape from such bad dictionaries. These replacement candidates are then further used to introduce a strategy for the automatic choice of the sparsity level S and the dictionary size K. Most of the material presented in this chapter and more can be found in [49].

4.1 Spurious Fixed Points

We now have a closer look at what happens when ITKrM does not find all atoms using a random initialisation. From [61] we know that ITKrM is most likely to not recover the full dictionary from a random initialisation when the signals are very sparse (S small) and the noiselevel is small. Since we want to closely inspect the resulting dictionaries, we only run a small experiment in $\mathbb{R}^{32}$, where we try to recover a very incoherent dictionary from 2-sparse vectors. The dictionary, containing 48 atoms, consists of the Dirac basis and the first half of the vectors from the Hadamard basis, and as such has coherence $\mu = 1/\sqrt{32} \approx 0.18$. The signals follow the model in (3.2), where the coefficient sequences c are constructed by choosing $b \in [0.9, 1]$ uniformly at random and setting $c_1 = 1/\sqrt{1+b^2}; c_2 = bc_1$ and $c_j = 0$ for $j \geq 3$. The noise is chosen to be Gaussian with variance $\rho^2 = 1/(16d)$, corresponding to SNR $= 16$. Running ITKrM with 20000 new signals per iteration for 25 iterations and 10 different random initialisations we recover 4 times 46 atoms and 6 times 44 atoms.

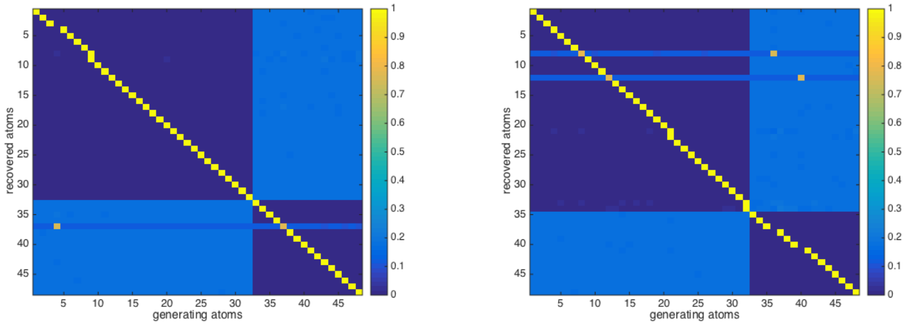


Figure 4.1: Cross-Gram matrices $\Psi^\star\Phi$ for recovered dictionaries with 2 (left) and 4 (right) missing atoms.

An immediate observation is that we always miss an even number of atoms. Taking a look at the recovered dictionaries - examples for recovery of 44 and 46 atoms are shown in Figure 4.1 - we see that this is due to their special structure. In case of $2n$ missing atoms, we always observe that n atoms of the generating dictionary are recovered twice and that n atoms in the learned dictionary are a 1:1 linear combinations of 2 missing atoms from the generating dictionary, respectively.
This shows that in the most simple case of 2 missing atoms (after rearranging and sign flipping the atoms in Φ) the recovered (and rearranged) dictionary Ψ has the form

$$\Psi = (\phi_1, \phi_1, \phi_3, \dots, \phi_{K-1}, \psi_K) \quad \text{with} \quad \psi_K = \frac{\phi_2 + \phi_K}{\sqrt{2 + 2\langle \phi_2, \phi_K \rangle}}. \tag{4.1}$$

Such a dictionary, or a slightly perturbed version of it, clearly cannot have the necessary cross-coherence property in Theorem 3.1 with any reasonably incoherent dictio-

nary Φ. In particular, we have that $\min_k |\langle\psi_k, \phi_k\rangle| \leq \mu(\Phi) \ll \gamma_{dyn} \cdot \log K \cdot \mu(\Phi, \Psi)$, which is quite contrary to (3.7).
To further see why ITKrM has problems to escape from such dictionaries, in the following, we have an even closer look at the special case of 2 missing atoms. In particular, we show that in case where the current dictionary estimate contains configurations as described above it is very likely that they are stable. For this, we first have to take a closer look at the individual components of ITKrM, meaning, thresholding as well as the residuals used for updating the dictionary atoms.

Thresholded support and approximate residual in case of 2 missing atoms

In case of 2 missing atoms, we now analyse which support thresholding will recover depending on whether the generating support contains the indices of the double or missing atoms, how the residuals $a = y - P(\Psi_{I^t})y$ for the current dictionary estimate Ψ will look like, as well as their corresponding probability.

We consider noiseless signals that are perfectly S-sparse in some dictionary Φ, $y = \Phi_I x_I = \Phi_I c_I \sigma_I$. For simplicity, let us assume that the non-zero coefficients c_i are equal to 1 and hence, $y = \Phi_I \sigma_I$. For the dictionary $\Psi = (\psi_1, \psi_2, \dots, \psi_K)$ obtained from ITKrM we assume that we have $\psi_1 = \psi_2 = \phi_1$, $\psi_i = \phi_i$ for all $i \in \{3, \dots, K-1\}$ and $\psi_K = (\phi_2 + h \cdot \phi_K)/\sqrt{2 + 2h\theta}$ with $h = 1$ if $\theta = \langle\phi_2, \phi_K\rangle \geq 0$ and $h = -1$ else. For our analysis we assume that w.l.o.g. $\langle\phi_2, \phi_K\rangle \geq 0$ and hence, $h = 1$. Further, we assume that each atom is equally likely to be picked.
For conciseness and in order to better deal with the recovered support sets we define for an index set I where the index $i \in I$ has been replaced by an index $j \notin I$, $I_{i\leftrightarrow j} := (I \setminus \{i\}) \cup \{j\}$. To estimate the probabilities of the residuals, we have to estimate the probability of the corresponding support I. In particular, to estimate the probability of a support I containing any $S - \ell$ indices from the set $\{3, \dots, K-1\}$ and ℓ specific indices $i \in \{1, 2, K\}$, we always use the formula $\binom{K-3}{S-\ell}\binom{K}{S}^{-1}$. For example, the probability that $I \subseteq \{3, \dots, K-1\}$, meaning $\ell = 0$, is $\binom{K-3}{S}\binom{K}{S}^{-1} \approx \left(1 - \frac{S}{K}\right)^3$.

In the following we estimate the thresholded supports, the residuals and their probability for all cases of generating supports $I \subseteq \{1, \dots, K\}$ with $|I| = S$. The results are then summarised in Table 4.1.

$\mathbf{1}, \mathbf{2}, \mathbf{K} \notin \mathbf{I}$: In this case, we have $I \subseteq \{3, 4, \dots, K-1\}$ with probability $\binom{K-3}{S}\binom{K}{S}^{-1} \approx \left(1 - \frac{S}{K}\right)^3$ and $\psi_i = \phi_i$ for all $i \in I$. For the inner products with the signal y, we have

$$\begin{aligned}
i \in I : |\langle \psi_i, \Phi_I \sigma_I \rangle| &= |\langle \phi_i, \phi_i \rangle \sigma_i + \langle \phi_i, \Phi_{I \setminus \{i\}} \sigma_{I \setminus \{i\}} \rangle| \\
&\geq 1 - (S-1)\mu \geq 1 - S\mu, \\
i \in I^c \setminus \{2, K\} : |\langle \psi_i, \Phi_I \sigma_I \rangle| &= |\langle \phi_i, \Phi_I \sigma_I \rangle| \leq S\mu, \\
i = 2 : |\langle \psi_2, \Phi_I \sigma_I \rangle| &= |\langle \phi_1, \Phi_I \sigma_I \rangle| \leq S\mu, \\
i = K : |\langle \psi_K, \Phi_I \sigma_I \rangle| &= |\langle (\phi_2 + \phi_K)/\sqrt{2 + 2\theta}, \Phi_I \sigma_I \rangle| \leq \sqrt{2} S\mu.
\end{aligned}$$

Hence, $I^t = I$ and

$$a = \Phi_I \sigma_I - P(\Psi_{I^t}) \Phi_I \sigma_I = \Phi_I \sigma_I - P(\Phi_I) \Phi_I \sigma_I = 0,$$

with probability approximately $\left(1 - \frac{S}{K}\right)^3$.

$\mathbf{1 \in I;\ 2, K \notin I:}$ For this case, for the probability of the support I we have $\binom{K-3}{S-1}\binom{K}{S}^{-1} \approx \frac{S}{K}\left(1 - \frac{S}{K}\right)^2$. For the inner products with y, we get

$$\begin{aligned}
i = 1 : |\langle \psi_1, \Phi_I \sigma_I \rangle| &= |\langle \phi_1, \Phi_I \sigma_I \rangle| \geq 1 - S\mu, \\
i = 2 : |\langle \psi_2, \Phi_I \sigma_I \rangle| &= |\langle \phi_1, \Phi_I \sigma_I \rangle| \geq 1 - S\mu, \\
i \in I \setminus \{1\} : |\langle \psi_i, \Phi_I \sigma_I \rangle| &= |\langle \phi_i, \Phi_I \sigma_I \rangle| \geq 1 - S\mu, \\
i \in I^c \setminus \{2, K\} : |\langle \psi_i, \Phi_I \sigma_I \rangle| &= |\langle \phi_i, \Phi_I \sigma_I \rangle| \leq S\mu, \\
i = K : |\langle \psi_K, \Phi_I \sigma_I \rangle| &= |\langle (\phi_2 + \phi_K)/\sqrt{2 + 2\theta}, \Phi_I \sigma_I \rangle| \leq \sqrt{2} S\mu,
\end{aligned}$$

and hence, $I^t \subseteq I \cup \{2\}$. Using that for $i \in I \setminus \{1\}$ we have $\operatorname{span}(\Psi_{(I \setminus \{i\}) \cup \{2\}}) = \operatorname{span}(\Phi_{I \setminus \{i\}})$, for the residual of $I^t = I_{i \leftrightarrow 2}$, we get

$$a = \Phi_I \sigma_I - P(\Psi_{I^t}) \Phi_I \sigma_I = \Phi_I \sigma_I - P(\Phi_{I \setminus \{i\}}) \Phi_I \sigma_I \approx \phi_i \sigma_i.$$

In case $I^t = I$ or $I^t = I_{1 \leftrightarrow 2}$ the residual is zero. Hence, the thresholded support and the residual take one of the following forms:

$$\begin{aligned}
&I^t = I_{i \leftrightarrow 2} \quad \text{for} \quad i \in I \setminus \{1\} \quad \text{with} \quad a \approx \phi_i \sigma_i, \\
&I^t = I \quad \text{with} \quad a = 0, \\
&I^t = I_{1 \leftrightarrow 2} \quad \text{with} \quad a = 0.
\end{aligned}$$

In order to estimate the probability of having the residual $\phi_i \sigma_i$, in addition to the probability of the support I, we have to take into account that with probability $\frac{S-1}{K-1}$ we have $i \in I \setminus \{1\}$. Hence, at worst with probability $\approx \frac{S^2}{K^2}\left(1 - \frac{S}{K}\right)^2$, we have $a \approx \phi_i \sigma_i$. Note that, on average this probability reduces to $\approx \frac{S}{K^2}\left(1 - \frac{S}{K}\right)^2$ as with probability $\frac{1}{S}$ the inner product with ψ_i is the smallest.

$\mathbf{2} \in \mathbf{I};\ \mathbf{1}, \mathbf{K} \notin \mathbf{I}:$ In this situation, we again have I with probability $\binom{K-3}{S-1}\binom{K}{S}^{-1} \approx \frac{S}{K}\left(1-\frac{S}{K}\right)^2$. Bounding the inner products with the signal, we obtain

$$\begin{aligned} i = 2 &: |\langle\psi_2, \Phi_I\sigma_I\rangle| = |\langle\phi_1, \Phi_I\sigma_I\rangle| \le S\mu, \\ i \in I\setminus\{2\} &: |\langle\psi_i, \Phi_I\sigma_I\rangle| = |\langle\phi_i, \Phi_I\sigma_I\rangle| \ge 1 - S\mu, \\ i = I^c\setminus\{K\} &: |\langle\psi_i, \Phi_I\sigma_I\rangle| = |\langle\phi_i, \Phi_I\sigma_I\rangle| \le S\mu \\ i = K &: |\langle\psi_K, \Phi_I\sigma_I\rangle| = |\langle(\phi_2+\phi_K)/\sqrt{2+2\theta}, \Phi_I\sigma_I\rangle| \ge (1-2S\mu)/\sqrt{2}. \end{aligned}$$

Therefore, $I^t = I_{2\leftrightarrow K}$ and for the corresponding residual, we get

$$\begin{aligned} a &= \Phi_I\sigma_I - P(\Psi_{I^t})\Phi_I\sigma_I = \phi_2\sigma_2 - P(\Psi_{I^t})\phi_2\sigma_2 \\ &= \left[\mathbb{I} - P(\Psi_{I^t})\right]\left(Q(\psi_K) + P(\psi_K)\right)\phi_2\sigma_2 \\ &= \left[\mathbb{I} - P(\Psi_{I^t})\right]Q(\psi_K)\phi_2\sigma_2 \approx (\phi_2 - \phi_K)\sigma_2/2, \end{aligned}$$

with probability $\approx \frac{S}{K}\left(1-\frac{S}{K}\right)^2$.

$\mathbf{K} \in \mathbf{I};\ \mathbf{1}, \mathbf{2} \notin \mathbf{I}:$ From the estimates of the inner products within the latter case, we see that we get $I^t = I$ and hence, for the residual

$$a = \Phi_I\sigma_I - P(\Psi_{I^t})\Phi_I\sigma_I \approx (\phi_K - P(\psi_K)\phi_K)\sigma_K \approx (\phi_K - \phi_2)\sigma_K/2,$$

with probability $\binom{K-3}{S-1}\binom{K}{S}^{-1} \approx \frac{S}{K}\left(1-\frac{S}{K}\right)^2$.

$\mathbf{1}, \mathbf{2} \in \mathbf{I};\ \mathbf{K} \notin \mathbf{I}:$ In this case, for the probability of the support I, we have $\binom{K-3}{S-2}\binom{K}{S}^{-1} \approx \frac{S^2}{K^2}\left(1-\frac{S}{K}\right)$. For the inner products with y, we get

$$\begin{aligned} i = 1, 2 &: |\langle\psi_i, \Phi_I\sigma_I\rangle| = |\langle\phi_1, \Phi_I\sigma_I\rangle| \ge 1 - S\mu, \\ i \in I\setminus\{1,2\} &: |\langle\psi_i, \Phi_I\sigma_I\rangle| = |\langle\phi_i, \Phi_I\sigma_I\rangle| \ge 1 - S\mu, \\ i \in I^c\setminus\{K\} &: |\langle\psi_i, \Phi_I\sigma_I\rangle| = |\langle\phi_i, \Phi_I\sigma_I\rangle| \le S\mu, \\ i = K &: |\langle\psi_K, \Phi_I\sigma_I\rangle| = |\langle(\phi_2+\phi_K)/\sqrt{2+2\theta}, \Phi_I\sigma_I\rangle| \ge (1-2S\mu)/\sqrt{2}, \end{aligned}$$

hence, $I^t = I$ and with probability $\approx \frac{S^2}{K^2}\left(1-\frac{S}{K}\right)$,

$$a = \Phi_I\sigma_I - P(\Psi_{I^t})\Phi_I\sigma_I = \Phi_I\sigma_I - P(\Phi_{I\setminus\{2\}})\Phi_I\sigma_I \approx \phi_2\sigma_2.$$

$\mathbf{1}, \mathbf{K} \in \mathbf{I};\ \mathbf{2} \notin \mathbf{I}:$ Similar to the latter case, for the inner products we have

$$\begin{aligned} i \in I\setminus\{K\} &: |\langle\psi_i, \Phi_I\sigma_I\rangle| = |\langle\phi_i, \Phi_I\sigma_I\rangle| \ge 1 - S\mu, \\ i \in I^c\setminus\{2\} &: |\langle\psi_i, \Phi_I\sigma_I\rangle| = |\langle\phi_i, \Phi_I\sigma_I\rangle| \le S\mu, \\ i = 2 &: |\langle\psi_2, \Phi_I\sigma_I\rangle| = |\langle\phi_1, \Phi_I\sigma_I\rangle| \ge 1 - S\mu, \\ i = K &: |\langle\psi_K, \Phi_I\sigma_I\rangle| = |\langle(\phi_2+\phi_K)/\sqrt{2+2\theta}, \Phi_I\sigma_I\rangle| \ge (1-2S\mu)/\sqrt{2}. \end{aligned}$$

Therefore, $I^t = I_{K\leftrightarrow 2}$ and

$$a = \Phi_I\sigma_I - P(\Psi_{I^t})\Phi_I\sigma_I = \Phi_I\sigma_I - P(\Phi_{I\setminus\{K\}})\Phi_I\sigma_I \approx \phi_K\sigma_K,$$

with probability $\binom{K-3}{S-2}\binom{K}{S}^{-1} \approx \frac{S^2}{K^2}\left(1-\frac{S}{K}\right)$.

$\mathbf{2, K \in I;\ 1 \notin I}$: In this situation, we again have I with probability $\binom{K-3}{S-2}\binom{K}{S}^{-1} \approx \frac{S^2}{K^2}\left(1-\frac{S}{K}\right)$ and for the inner products with the signals, we get

$$\begin{aligned}
i = 2 &: |\langle\psi_2, \Phi_I\sigma_I\rangle| = |\langle\phi_1, \Phi_I\sigma_I\rangle| \leq S\mu,\\
i \in I\setminus\{2,K\} &: |\langle\psi_i, \Phi_I\sigma_I\rangle| = |\langle\phi_i, \Phi_I\sigma_I\rangle| \geq 1 - S\mu,\\
i \in I^c &: |\langle\psi_i, \Phi_I\sigma_I\rangle| = |\langle\phi_i, \Phi_I\sigma_I\rangle| \leq S\mu,\\
i = K &: |\langle\psi_K, \Phi_I\sigma_I\rangle| = |\langle(\phi_2+\phi_K)/\sqrt{2+2\theta}, \Phi_I\sigma_I\rangle|\\
&\qquad = |\langle(\phi_2+\phi_K)/\sqrt{2+2\theta}, \phi_2\sigma_2 + \phi_K\sigma_K\rangle\\
&\qquad\quad + \langle(\phi_2+\phi_K)/\sqrt{2+2\theta}, \Phi_{I\setminus\{2,K\}}\sigma_{I\setminus\{2,K\}}\rangle|.
\end{aligned}$$

In order to get a bound for the inner product with ψ_K we have to distinguish whether the signal coefficients have the same sign σ_2, σ_K or not. If $\sigma_2 = \sigma_K$, we have $|\langle\psi_K, \Phi_I\sigma_I\rangle| \geq \sqrt{2}(1 - S\mu)$, thus $K \in I^t$ and for $I^t = I \setminus \{2\}$ we get

$$a = \Phi_I\sigma_I - P(\Psi_{I^t})\Phi_I\sigma_I = (\mathbb{I}_d - P(\psi_K))(\phi_2+\phi_K)\sigma_2 = 0.$$

To have $|I^t| = S$, we have to add another index $j \in I^c \cup \{2\}$, hence yielding $I^t = I$ or $I^t = I_{2\leftrightarrow j}$.
Conversely, if $\sigma_2 \neq \sigma_K$, the contribution of ϕ_2 and ϕ_K to the signal is orthogonal to ψ_K and hence, $|\langle\psi_K, \Phi_I\sigma_I\rangle| \leq \sqrt{2}(S-2)\mu$. From this we see that for small sparsity levels ($S \leq 6$) it is very unlikely that K will be contained in I^t and instead two indices $j, k \in I^c \cup \{2\}$ have to be added, likely those which are most correlated with the residual

$$a = \Phi_I\sigma_I - P(\Psi_{I^t})\Phi_I\sigma_I \approx \Phi_I\sigma_I - P(\Phi_{I\setminus\{2,K\}})\Phi_I\sigma_I \approx \pm(\phi_2 - \phi_K).$$

Hence, in case 2, $K \in I$ and $1 \notin I$, the thresholded support and the residuals take one of the following forms:

$$\begin{aligned}
&I^t = I_{2\leftrightarrow j} \quad \text{for} \quad j \in I^c \cup \{2\} \quad \text{with} \quad a = 0,\\
&I^t = I_{\{2,K\}\leftrightarrow\{j,k\}} \quad \text{for} \quad j,k \in I^c \cup \{2\} \quad \text{with} \quad a \approx \pm(\phi_2 - \phi_K).
\end{aligned}$$

In order to estimate the probability of the residuals, in addition to the probability of the support I, we have to take into account that with probability $\frac{1}{2}$ we have $\sigma_2 = \sigma_K$ resp. $\sigma_2 \neq \sigma_K$.

$\mathbf{1}, \mathbf{2}, \mathbf{K} \in \mathbf{I}$: For the probability of the support I, we have $\binom{K-3}{S-3}\binom{K}{S}^{-1} \approx \frac{S^3}{K^3}$. This case works analogue to the latter one, taking into account that $2 \in I^t$. In particular, for the inner products we have

$$\begin{aligned} i \in I \setminus \{K\} &: |\langle \psi_i, \Phi_I \sigma_I \rangle| \geq 1 - S\mu, \\ i \in I^c &: |\langle \psi_i, \Phi_I \sigma_I \rangle| \leq S\mu, \\ i = K &: |\langle \psi_K, \Phi_I \sigma_I \rangle| = |\langle (\phi_2 + \phi_K)/\sqrt{2+2\theta}, \Phi_I \sigma_I \rangle|. \end{aligned}$$

For the inner product with ψ_K, we again have to distinguish whether the signs σ_2 and σ_K are the same or not. Hence, we get

$$\begin{aligned} &I^t = I \quad \text{with} \quad a = 0 \quad \text{or} \\ &I^t = I_{K \leftrightarrow j} \quad \text{for} \quad j \in I^c \quad \text{with} \quad a \approx \pm(\phi_2 - \phi_K). \end{aligned}$$

The estimates obtained for the various cases are summarised in Table 4.1.

support I with $\lvert I \rvert = S$	**thresholded support**	**approximate residual**	**approx. prob. of residual**
$\{\mathbf{1}, \mathbf{2}, \mathbf{K}\} \cap \mathbf{I} = \emptyset$	$I^t = I$	0	$\left(1 - \frac{S}{K}\right)^3$
$\mathbf{1} \in \mathbf{I}, \{\mathbf{2}, \mathbf{K}\} \cap \mathbf{I} = \emptyset$	$I^t = I_{i \leftrightarrow 2}$, $i \in I \setminus \{1\}$	$\pm\phi_i$	$\lesssim \frac{S^2}{K^2}\left(1 - \frac{S}{K}\right)^2$
	$I^t = I$	0	$\lesssim \frac{S}{2K}\left(1 - \frac{S}{K}\right)^2$
	$I^t = I_{1 \leftrightarrow 2}$	0	$\lesssim \frac{S}{2K}\left(1 - \frac{S}{K}\right)^2$
$\mathbf{2} \in \mathbf{I}, \{\mathbf{1}, \mathbf{K}\} \cap \mathbf{I} = \emptyset$	$I^t = I_{2 \leftrightarrow K}$	$\pm(\phi_2 - \phi_K)/2$	$\frac{S}{K}\left(1 - \frac{S}{K}\right)^2$
$\mathbf{K} \in \mathbf{I}, \{\mathbf{1}, \mathbf{2}\} \cap \mathbf{I} = \emptyset$	$I^t = I$	$\pm(\phi_2 - \phi_K)/2$	$\frac{S}{K}\left(1 - \frac{S}{K}\right)^2$
$\{\mathbf{1}, \mathbf{2}\} \subseteq \mathbf{I}, \mathbf{K} \notin \mathbf{I}$	$I^t = I$	$\pm\phi_2$	$\frac{S^2}{K^2}\left(1 - \frac{S}{K}\right)$
$\{\mathbf{1}, \mathbf{K}\} \subseteq \mathbf{I}, \mathbf{2} \notin \mathbf{I}$	$I^t = I_{K \leftrightarrow 2}$	$\pm\phi_K$	$\frac{S^2}{K^2}\left(1 - \frac{S}{K}\right)$
$\{\mathbf{2}, \mathbf{K}\} \subseteq \mathbf{I}, \mathbf{1} \notin \mathbf{I}$	$I^t = I_{2 \leftrightarrow j}$, $j \in I^c \cup \{2\}$	0	$\frac{S^2}{2K^2}\left(1 - \frac{S}{K}\right)$
	$I^t = I_{\{2,K\} \leftrightarrow \{i,j\}}$, $i, j \in I^c \cup \{2\}$	$\pm(\phi_2 - \phi_K)$	$\frac{S^2}{2K^2}\left(1 - \frac{S}{K}\right)$
$\{\mathbf{1}, \mathbf{2}, \mathbf{K}\} \subseteq \mathbf{I}$	$I^t = I$	0	$\frac{S^3}{2K^3}$
	$I^t = I_{K \leftrightarrow j}$, $j \in I^c$	$\pm(\phi_2 - \phi_K)$	$\frac{S^3}{2K^3}$

Table 4.1: Thresholded support with corresponding approximate residual and the probability of having this residual, for various signal generating supports I.

From Table 4.1, we see that thresholding will correctly identify all supports which do not contain 1, 2 or K. This means that all atoms $\psi_3, \ldots, \psi_{K-1}$ will hardly be affected by the error originating from the failure of thresholding and hence, stay close to $\phi_3, \ldots, \phi_{K-1}$. For supports containing 1 and not 2 we have that they are rather unlikely to be identified correctly. In this case we will usually recover 1 and 2 and miss some other atom ϕ_i with $i \in I \setminus \{1\}$. Considering the residuals and their corresponding probability, this however will only rarely affect the atom update and hence, ψ_1 and ψ_2 will remain close to ϕ_1. From all the other cases where thresholding has failed we see that the most likely non-zero residual is given by $\pm(\phi_2 - \phi_K)$, whereas the probability of having $\pm\phi_2$ or $\pm\phi_K$ is much lower. In consequence, when updating ψ_2 and ψ_K, we quite rarely (ψ_2) or even never (ψ_K) have residuals which are purely the desired ones, i.e. ϕ_2 and ϕ_K, respectively, and hence, it is very unlikely that they will be drawn into the direction we want to have. As the $1:1$ combination ψ_K is the best possible approximation to both ϕ_2 and ϕ_K, we have that ψ_K will stay where it is. This means, in case of dictionaries Ψ with one double atom and one atom which corresponds to a $1:1$ combination as described above, one iteration of ITKrM will stay close to Ψ.

From the estimates above we can see that with probability $\left(1-\frac{S}{K}\right)^2$ the residual is zero (or close to zero in case of noise), with probability at most $\frac{S^2}{K^2}\left(1-\frac{S}{K}\right)$ it is close to ϕ_i for some i, and with probability at least $\frac{2S}{K}\left(1-\frac{S}{K}\right)^2$ it is close to $\pm(\phi_2-\phi_K)$. Hence, the most likely non-zero residual is a linear combination of the two missing atoms ϕ_2 and ϕ_K, or to be more precise, the residuals are very likely to be 1-sparse in the complementary $1:1$ combinations $\pm(\phi_2-\phi_K)$. In order to get an idea of how the residuals look like in case of more than 2 missing atoms, in the following, we briefly discuss the general case.

Approximate residual and its probability in case of $2n$ missing atoms

In case of $2n$ missing atoms, we now analyse how the residuals $a = y - P(\Psi_{I^t})y$ will look like, for which kind of generating support this can happen, as well as the probability of their occurrence.

Similar to the previous results we consider noiseless signals that are perfectly S-sparse in some dictionary Φ, $y = \Phi_I x_I = \Phi_I c_I \sigma_I$. For simplicity, we again assume that the non-zero coefficients c_i are equal to 1 and hence, $y = \Phi_I \sigma_I$. Further, we assume that each atom is equally likely to be picked.
For the dictionary $\Psi = (\psi_1, \psi_2, \dots, \psi_K)$ obtained from ITKrM, we assume that we have $\psi_{v_i} = \psi_{\bar{v}_i} = \phi_{v_i}$, $\psi_{w_i} = (\phi_{\bar{v}_i} + h_i \cdot \phi_{w_i})/\sqrt{2+2h_i\theta_i}$, with $h_i = 1$ if $\theta_i = \langle \phi_{\bar{v}_i}, \phi_{w_i}\rangle \geq 0$ and $h_i = -1$ else, for all $i \in \{1,\dots,n\}$ and $\psi_{u_i} = \phi_{u_i}$ for all $i \in \{1,\dots,K-3n\}$. W.l.o.g. we assume that for all i we have $\langle \phi_{\bar{v}_i}, \phi_{w_i}\rangle \geq 0$ and therefore $h_i = 1$. Further, let us define the sets $V = (v_1,\dots,v_n)$ and $\bar{V} = (\bar{v}_1,\dots,\bar{v}_n)$ containing the indices of the double atoms, $W = (w_1,\dots,w_n)$ the set consisting of the indices corresponding to the $1:1$ combinations and $U = \{1,\dots,K\} \setminus (V \cup \bar{V} \cup W) = (u_1,\dots,u_{K-3n})$ the set of indices of the single atoms.
In order to estimate the probability of the residuals we again have to estimate the probability of the corresponding support I. Let $\ell = \ell_V + \ell_{\bar{V}} + \ell_W$ denote the number of special atoms in I, meaning the ones with indices in $V \cup \bar{V} \cup W$ such that $|I \cap V| = \ell_V$, $|I \cap \bar{V}| = \ell_{\bar{V}}$ and $|I \cap W| = \ell_W$. To estimate the probability of supports containing any ℓ_V indices from V, any $\ell_{\bar{V}}$ indices from $\bar{V}$, any ℓ_W indices from W and

any $S - \ell$ indices from U, we always use the formula

$$\binom{n}{\ell_V}\binom{n}{\ell_{\bar{V}}}\binom{n}{\ell_W}\binom{K-3n}{S-\ell}\binom{K}{S}^{-1}.$$

On the other hand, to estimate the probability of supports I containing ℓ_V specific indices $v_i \in V$ and any $\ell_{\bar{V}}$ indices from $\bar{V}$, any ℓ_W indices fom W and any $S-\ell$ indices from U, we use

$$\binom{n}{\ell_{\bar{V}}}\binom{n}{\ell_W}\binom{K-3n}{S-\ell}\binom{K}{S}^{-1}.$$

Similarly, if we estimate the probability of supports I containing $\ell_{\bar{V}}$ specific indices from $\bar{V}$ or ℓ_W specific indices from W and any other indices from the remaining sets. To estimate the probability of supports I containing ℓ_U specific indices and any $S-\ell_U-\ell$ indices from U together with any ℓ_V, $\ell_{\bar{V}}$ and ℓ_W indices from V, $\bar{V}$ and W, respectively, we use the formula

$$\binom{n}{\ell_V}\binom{n}{\ell_{\bar{V}}}\binom{n}{\ell_W}\binom{K-\ell_U-3n}{S-\ell_U-\ell}\binom{K}{S}^{-1}.$$

For example, for the probability of $I \subseteq U$ and hence $\ell_V = \ell_{\bar{V}} = \ell_W = 0$, we have $\binom{K-3n}{S}\binom{K}{S}^{-1} \approx \left(1-\frac{S}{K}\right)^{3n}$.

The following table lists the estimates for the most common cases, with $q := \left(1-\frac{S}{K}\right)$.

support I with $\lvert I\rvert = S$	**approximate residual**	**approx. prob. of residual**
$I \subseteq U$	0	$\left(1-\frac{S}{K}\right)\cdot q^{3n-1}$
$\lvert I\cap U\rvert = S-1$, $\lvert I\cap V\rvert = 1$ and $u_k \in I\cap U$	0	$\frac{nS}{K}\cdot q^{3n-1}$
	$\pm\phi_{u_k}$	$\lesssim \frac{nS^2}{K^2}\cdot q^{3n-1}$
$\lvert I\cap U\rvert = S-1$ and $I\cap\bar{V} = \bar{v}_i$	$\pm(\phi_{\bar{v}_i}-\phi_{w_i})/2$	$\frac{S}{K}\cdot q^{3n-1}$
$\lvert I\cap U\rvert = S-1$ and $I\cap W = w_i$	$\pm(\phi_{\bar{v}_i}-\phi_{w_i})/2$	$\frac{S}{K}\cdot q^{3n-1}$
$\lvert I\cap U\rvert = S-2$, $\lvert I\cap V\rvert = 1$ and $I\cap\bar{V} = \bar{v}_i$	$\pm\phi_{\bar{v}_i}$	$\frac{nS^2}{K^2}\cdot q^{3n-2}$
$\lvert I\cap U\rvert = S-2$, $\lvert I\cap V\rvert = 1$ and $I\cap W = w_i$	$\pm\phi_{w_i}$	$\frac{nS^2}{K^2}\cdot q^{3n-2}$
$\lvert I\cap U\rvert = S-2$, $\lvert I\cap V\rvert = 2$ and $u_k, u_\ell \in I\cap U$	0	$\frac{n^2S^2}{K^2}\cdot q^{3n-2}$
	$\pm\phi_{u_k}$	$\lesssim \frac{n^2S^3}{K^3}\cdot q^{3n-2}$
	$\pm(\phi_{u_k}\pm\phi_{u_\ell})$	$\lesssim \frac{n^2S^4}{K^4}\cdot q^{3n-2}$

Table 4.2: Approximate residual and the probability of having this residual in case of $2n$ missing atoms, for various signal generating supports I.

From the results in Table 4.2 we see that the most likely non-zero residuals are linear combinations of 2 missing atoms. In particular, they are complements of the

$1:1$ combinations. We also see that the probability of having residuals which are purely the missing ones is much lower but increases with the number of double atoms. Similarly, the average occurrence of a residual $\pm\phi_{u_k}$ increases with n to $\frac{nS}{K^2}$, which is still quite rare. In summary this shows that also in case of $2n$ missing atoms, the residuals tend to be 1-sparse in the complements of the $1:1$ combinations ψ_{w_i}. The estimates of the residuals in Table 4.2 were obtained using similar considerations as in the previous results.

Summarising the results found so far, in the previous chapter we have seen that ITKrM may not be a contraction if the current dictionary estimate is too coherent, has large operator norm or if two estimated atoms are close to one generating atom. In this section, we have also seen that whenever ITKrM does not recover the generating dictionary it produces a dictionary without the necessary cross-coherence property from Theorem 3.1. In particular, the resulting dictionaries most likely contain atoms of some special structure which clearly cannot satisfy this condition. Moreover, in cases where the current dictionary estimate contains such configurations of generating atoms, they are also very likely to be stable. To overcome this problem and therefore help ITKrM to escape from such bad dictionaries, we have to introduce some new ideas. For that, in the next section, we develop a replacement strategy where we use the information that these stable fixed points contain several double atoms (coherent atoms) and that the residuals contain information about the missing atoms (they are 1-sparse in the complements of the $1:1$ combinations of the missing atoms).

4.2 Replacement

In the previous section we have seen that whenever ITKrM does not recover the generating dictionary it produces a dictionary which can be characterised as very coherent. Replacement of coherent atoms with new, randomly drawn atoms is a simple clean-up step that most dictionary learning algorithms based on alternating minimisation, e.g. K-SVD, employ additionally in each iteration. While randomly drawing a replacement candidate is cost-efficient and democratic, the drawback is that the new atom converges only very slowly or not at all to the missing generating atom.
To see why a randomly drawn replacement atom is not the best idea, let us again consider the case where the current dictionary estimate Ψ contains one double atom $\psi_1 = \psi_2 = \phi_1$ and one $1:1$ atom $\psi_K \propto \phi_2 + \phi_K$.

For a replacement atom ψ_{new} drawn uniformly at random from the d-dimensional unit sphere, we have for any fixed vector v, $\|v\|_2 = 1$,

$$\mathbb{P}(|\langle v, \psi_{\text{new}}\rangle| \geq t) \leq 2\exp\left(-\frac{t^2 d}{2}\right),$$

and hence, for any atom ϕ_k, $|\langle \phi_k, \psi_{\text{new}}\rangle| \lesssim \sqrt{2\log(2K)/d}$. This means, replacing ψ_2 with ψ_{new}, with high probability ψ_{new} will be quite incoherent to all atoms ϕ_k, meaning $|\langle \phi_k, \psi_{\text{new}}\rangle| \ll |\langle \phi_k, \psi_k\rangle|$ for $k \neq 2$ and $|\langle \phi_2, \psi_{\text{new}}\rangle| \ll |\langle \phi_2, \psi_K\rangle|$, and so never be picked. Hence, we would replace a coherent atom with an unused atom. In particular, going back to the thresholding analysis in Section 4.1, we see that the only time it has a chance to be picked is if $I \cap \{1, 2, K\} = \{2, K\}$. Here we distinguished between two cases, $\sigma_2 = \sigma_K$ and $\sigma_2 = -\sigma_K$. In case $\sigma_2 = \sigma_K$ the residual is zero and so even if ψ_{new} is picked it will not be drawn into a useful direction. This means, the only useful case where ψ_{new} has a chance to be picked is if $I \cap \{1, 2, K\} = \{2, K\}$ and the signal contains the rare constellation $\phi_2 - \phi_K$. Unfortunately, with high probability, ψ_{new} is also incoherent to this linear combination and so might actually never be picked. However, looking on the bright side, we also see that once it is picked, the updated atom $\bar{\psi}_2$ will be very close to $\phi_2 - \phi_K$ since we have $y - P(\Psi_{I^t})y \approx (\phi_2 - \phi_K)$ and $|\langle y, \psi_{\text{new}}\rangle| \lesssim \sqrt{2\log(2K)/d}$. Thus, in the next iteration, the updated atom $\bar{\psi}_2 \approx (\phi_2 - \phi_K)/\sqrt{2 - 2\theta}$ will be serious competition for $\bar{\psi}_K \approx (\phi_2 + \phi_K)/\sqrt{2 + 2\theta}$ in the thresholding of all signals containing either ϕ_2 or ϕ_K. This iteration will then create a first imbalance of the ratio between ϕ_2 and ϕ_K within one or both of the estimated atoms, making one the more likely choice for ϕ_2 and the other the more likely choice for ϕ_K in the subsequent iteration. There the imbalance will be further increased until a few iterations later we finally have $\psi_2 \approx \phi_2$ and $\psi_K \approx \phi_K$ or the other way around. However, from this we can see that the chances of becoming correlated to $\phi_2 - \phi_K$ are very low. Thus the natural next question is whether we can do better than a random replacement. To find a smarter strategy we use the insights gained in Section 4.1.

4.2.1 Learning from bad dictionaries

Looking back to the thresholding analysis in Section 4.1, we have seen that in case of $2n$ missing atoms the most likely non-zero residuals are linear combinations of the missing atoms. To be more precise, the residuals tend to be 1-sparse in the $1:1$ complements $(\phi_{\bar{v}_i} - h_i\phi_{w_i})$ of the $1:1$ combinations. The idea is now that we learn these $1:1$ complements and in each iteration of ITKrM additionally employ a clean-up step where we replace coherent atoms with these learned atoms. As we have argued above, these will be serious competition for the 1:1 combinations and quickly rotate into the correct configuration.

Knowing that the residuals are very likely to be 1-sparse in the complements of the $1:1$ combinations, $(\phi_{\bar{v}_i} - h_i\phi_{w_i})$, they can be simply obtained by running ITKrM, which for $S = 1$ reduces to ITKsM, on the residuals. Concretely, we choose the number

$L \ll K$ of candidate atoms, meaning the maximal number of atoms we can replace after each iteration, initialise a $d \times L$ dictionary $\Gamma = (\gamma_1 \dots \gamma_L)$ of candidates and in each iteration of ITKrM add the following clean-up steps. For all signals we find $i_n = \arg\max_\ell |\langle \gamma_\ell, a_n \rangle|$, where $a_n = y_n - P(\Psi_{I_n^t}) y_n$ and update the candidate atoms as $\bar{\gamma}_\ell = \sum_{n: i_n = \ell} a_n \cdot \text{sign}(\langle \gamma_\ell, a_n \rangle)$ with subsequent normalisation.

We can also immediately see the advantages of this strategy over other residual based replacement strategies, such as using the largest principal components or the largest residuals, [55, 30]. In the case of noise or outliers, the largest residuals are most likely to be outliers or pure noise, meaning that this strategy effectively corresponds to random replacement. The largest principal components of the residuals on the other hand, will be a balanced linear combinations of several correlated 1:1 complementary atoms, and as such less serious competition for the original 1:1 combinations during thresholding.
After learning enough from bad dictionaries to inspire a promising replacement strategy, the next subsection will deal with its practical implementation.

4.2.2 Replacement in detail

Now that we have laid out the basic strategy, it remains to deal with all the details. For instance, if we have used all replacement candidates after one iteration, after the next iteration the replacement candidates might not have converged yet.

Efficient learning of replacement atoms

To solve the above mentioned problem, observe that the number of replacement candidates will be much smaller than the dictionary size, $L \ll K$. Therefore, we need less training signals per iteration to learn the candidates or equivalently we can update Γ more frequently, meaning, we renormalise after each batch of $N_\Gamma < N$ signals and set $\Gamma = \bar{\Gamma}$. Like this, every augmented iteration of ITKrM will produce L replacement candidates.

Combining coherent atoms

The next questions concern the actual replacement procedure. Assume we have fixed a threshold $\mu_{\max}$ for the maximal coherence. If our estimate Ψ contains two atoms whose mutual coherence is above the threshold, $|\langle \psi_k, \psi_{k'} \rangle| > \mu_{\max}$, which atom should be replaced? One strategy that has been employed for instance in the context of analysis operator learning, [17], is to average the two atoms, that is, to set $\psi_k^{\text{new}} = \psi_k + \text{sign}(\langle \psi_k, \psi_{k'} \rangle) \psi_{k'}$. The reasoning is that if both atoms are good approximations to the generating atom ϕ_k, then their average will be an even better approximation. However, if one atom ψ_k is already a very good approximation to the generating atom $\psi_k \approx \phi_k$ while $\psi_{k'}$ is still as far away as indicated by $\mu_{\max}$, that is $\psi_{k'} \approx \mu_{\max}\phi_k + \sqrt{1-\mu_{\max}^2}z_k$, then the averaged atom will be a worse approximation than ψ_k and it would be preferable to simply keep ψ_k.

To determine which of two coherent atoms is the better approximation, we note that the better approximation to ϕ_k should be more likely to be selected during thresholding. This means that we can simply count how often each atom is contained in the thresholded supports I_n^t, $v(k) = \sharp\{n : k \in I_n^t\}$, and in case of two coherent atoms keep the more frequently used one. Based on the value function v we can also employ a weighted merging strategy and set $\psi_k^{\text{new}} = v(k)\psi_k + \operatorname{sign}(\langle\psi_k, \psi_{k'}\rangle)v(k')\psi_{k'}$. If both atoms are equally good approximations, then their value functions should be similar and the balanced combination will be a better approximation. If one atom is a much better approximation it will be used much more often and the merged atom will correspond to this better atom.

Selecting a candidate atom

Having chosen how to combine two coherent atoms, we next need to decide which of our L replacement candidates we are going to use. To keep the dictionary incoherent, we first discard all candidates γ_ℓ, whose maximal coherence with the remaining dictionary atoms is larger than our threshold, that is, $\max_k |\langle\gamma_\ell, \phi_k\rangle| \geq \mu_{\max}$.

To decide which remaining candidate is likely to be the most valuable, we use a counter similar to the one for the dictionary atoms. However, we have to be more careful here since every residual is added to one candidate. If the residual contains only noise, which happens in most cases, and the candidates are reasonably incoherent to each other, then each candidate is equally likely to have its counter increased. This means that the candidate atom that actually encodes the missing atom (or 1:1 complement) will only be slightly more often used than the other candidates. So to better distinguish between good and bad candidates, we additionally employ a threshold τ and set $v_\Gamma(\ell) = \sharp\{n : \ell = i_n, |\langle\gamma_\ell, a_n\rangle| \geq \tau\|a_n\|\}$. To determine the size of the threshold, observe that for a residual consisting only of Gaussian noise, $a = r$, we have for any γ_ℓ the bound

$$\mathbb{P}(|\langle\gamma_\ell, r\rangle| \geq \tau\|r\|_2) \leq 2\exp\left(-\frac{d\tau^2}{2}\right), \tag{4.2}$$

which for $\tau = \sqrt{2\log(2K)/d}$ becomes $1/K$. This means that the contribution to $v_\Gamma(\ell)$ from all the pure noise residuals is at best N/K. On the other hand, with probability S/K, the residual will encode the missing atom or 1:1 complement $a \approx (\phi_i - \phi_j)\cdot|x_i|/2$. For reasonable sparsity levels, $S \lesssim \frac{d}{4\log(2K)}$, and signal to noise ratios, the candidate γ_ℓ closest to the missing atom will be picked and should have inner product of the size $|\langle \gamma_\ell, a\rangle| \approx |x_i|/2 \approx \frac{1}{2\sqrt{S}} \gtrsim \tau \|a\|_2$. This means that for a good candidate the value function will be closer to NS/K.

Dealing with unused atoms

If an atom has never been updated, or more generally, if the norm of the new estimator is too small, we simply do not update this atom but set the associated value function to zero. After replacing all coherent atoms we then proceed to replace these unused atoms.

The combination of all these considerations leads to an augmented version of the ITKrM algorithm. More details, a summary of the algorithm as well as numerical experiments testing its performance on synthetic data can be found in the original paper [49]. Here we next address the question of how to learn dictionaries without the knowledge of the correct sparsity level S and dictionary size K, leading to an adaptive algorithm whose performance on real data we investigate in Chapter 5.

4.3 Adaptive Dictionary Learning

While a replacement strategy as described above improves the global convergence behaviour of ITKrM, its performance also strongly depends on the choice of the sparsity level S and the dictionary size K which are needed to be given as input parameter. In this section, we discuss the difficulties which arise when choosing S and K and develop a strategy for their adaptive choice.
We first investigate how to adaptively choose the sparsity level S for a dictionary of fixed size K.

4.3.1 Adapting the sparsity level

The choice of the sparsity level S is very difficult as (like K) it influences both the convergence speed and the final precision of the learned dictionary. In the following, let S denote the correct sparsity level, meaning, the sparsity level of the signals, and S_e the sparsity level given to the algorithm. We first have a closer look at the advantages and drawbacks of both under- and overestimating the sparsity level.
When underestimating the sparsity level, meaning providing $S_e < S$ instead of S, we know from the numerical experiments in [49] that the algorithm tends to recover the generating dictionary in less iterations than with the true sparsity level. To be more precise, the computational complexity of an iteration increases with S_e, so a

smaller sparsity level leads to faster convergence not only in terms of iterations but also reduces the computation time per iteration. The advantage of overestimating the sparsity level, $S_e > S$ on the other hand, is the potentially higher precision, so the final error between the recovered and the generating dictionary (atoms), can be smaller than for the true sparsity level S. Intuitively, this is due to the fact that for $S_e > S$, thresholding with the generating dictionary is more likely to recover the correct support, in the sense that $I \subset I^t$. For a clean signal, $y = \Phi_I x_I$ this means that the residual is zero and hence, the estimate of every atom ϕ_i with $i \in I^t$, even if $i \notin I$, is simply reinforced by itself $\langle \phi_i, y \rangle \phi_i$. However, in a noisy situation, $y = \Phi_I x_I + r$, where the residual has the shape $a = Q(\Phi_{I^t})r$, the estimate of the additional atom $i \in I^t/I$ is not only reinforced but also disturbed by adding noise in form of the residual once more than necessary.
To further see why both under- and overestimating the sparsity level comes with risks, assume that we allow $S + 1$ instead of the true sparsity level S, for perfectly sparse and clean signals. In this case, any dictionary derived from the generating dictionary by replacing a pair of atoms (ϕ_i, ϕ_j) by $(\tilde{\phi}_i, \tilde{\phi}_j) = A(\phi_i, \phi_j)$ for an invertible (well conditioned) matrix A, will provide perfectly $S+1$-sparse representations to the signals and be a fixed point of ITKrM. On the other hand, providing $S - 1$ instead of S can have even more dire consequences since we can replace any generating atom with a random vector and again have a fixed point of ITKrM. If the original dictionary is an orthonormal basis and the sparse coefficients have equal size in absolute value, any such disturbed estimator even gives the same approximation quality. However, in more realistic scenarios, where we have coherence, noise or imbalanced coefficients and therefore the missing atom has the same probability as the others to be among the $S - 1$ atoms most contributing to a signal, the generating dictionary should still provide the smallest average approximation error.

Adaptive choice of the sparsity level S

The above considerations show that the choice of the sparsity level S is accompanied by many difficulties. The idea is now that whenever we have coherence, noise or imbalanced coefficients, the signals can be interpreted as being 1-sparse (with enormous error and miniscule gap $c(1)/c(2)$) in the generating dictionary. This means, learning with $S_e = 1$ should lead to a reasonable first estimate of most atoms. Of course if the signals are not actually 1-sparse this estimate will be somewhere between rough, for small S, and unrecognisable, for larger S, and the question is how to decide whether we should increase S_e. If we already had the generating dictionary, the simplest way would be to look at the residuals and see how much we can decrease their energy by adding another atom to the support. A lower bound for the decrease of a residual a can be simply estimated by calculating $\max_k(\langle \phi_k, a \rangle)^2$.
If we have the correct sparsity level and thresholding recovers the correct support $I^t = I$, the residual consists only of noise, $a = Q(\Phi_I)(\Phi_I x_I + r) = Q(\Phi_I)r \approx r$. For a

Gaussian noise vector r and a given threshold $\theta \cdot \|r\|_2$, we now estimate how many of the remaining $K - S$ atoms can be expected to have inner products larger than $\theta \cdot \|r\|_2$ as

$$\begin{aligned}\mathbb{E}\left(\sharp\{k : |\langle r, \phi_k\rangle|^2 > \theta^2 \cdot \|r\|_2^2\}\right) &= \sum_k \mathbb{P}\left(|\langle r, \phi_k\rangle|^2 > \theta^2 \cdot \|r\|_2^2\right) \\ &< 2(K-S)\exp\left(-\frac{d\theta^2}{2}\right). \end{aligned} \tag{4.3}$$

In particular, setting $\theta = \theta_K := \sqrt{2\log(4K)/d}$ the expectation above is smaller than $\frac{1}{2}$. This means that if we take the empirical estimator of the expectation above, using the approximation $r_n \approx a_n$, we should get

$$\frac{1}{N}\sum_n \sharp\{k : |\langle a_n, \phi_k\rangle|^2 > \theta_K^2 \cdot \|a_n\|_2^2\} \lesssim \frac{1}{2}, \tag{4.4}$$

which rounds to zero, indicating that we have the correct sparsity level.
Conversely, if we underestimate the correct sparsity level, $S_e = S - m$ for $m > 0$, then thresholding can necessarily only recover part of the correct support, $I^t \subset I$. Denote the set of missing atoms by $I^m = I/I^t$. The residual has the shape

$$a = Q(\Phi_{I^t})(\Phi_I x_I + r) = Q(\Phi_{I^t})(\Phi_{I^m} x_{I^m} + r) \approx \Phi_{I^m} x_{I^m} + r$$

For all missing atoms $i \in I^m$ the squared inner products are approximately

$$|\langle a, \phi_i\rangle|^2 \approx (x_i + \langle r, \phi_i\rangle)^2.$$

Assuming well-balanced coefficients, where $|x_i| \approx 1/\sqrt{S}$ and therefore $\|\Phi_{I^m} x_{I^m}\|_2^2 \approx m/S$, a sparsity level $S \lesssim \frac{d}{2\log(4K)}$ and reasonable noiselevels, this means that with probability at least $\frac{1}{2}$, we have for all $i \in I^m$

$$|\langle a, \phi_i\rangle|^2 \gtrsim |x_i|^2 \gtrsim \frac{1}{2m}(\|\Phi_{I^m} x_{I^m}\|_2^2 + \|r\|_2^2) \gtrsim \theta_K^2 \|a\|_2^2,$$

and in consequence

$$\frac{1}{N}\sum_n \sharp\{k : |\langle a_n, \phi_k\rangle|^2 > \theta_K^2 \cdot \|a_n\|_2^2\} \gtrsim \frac{m}{2}. \tag{4.5}$$

This rounds to at least 1, indicating that we should increase the sparsity level.

Based on the two estimates above and starting with sparsity level $S_e = 1$, we should be able to arrive at the correct sparsity level S. Unfortunately, the indicated update rule for the sparsity level is too simplistic in practice as it relies on thresholding always finding the correct support, given the correct sparsity level.

Assume that $S_e = S$ but thresholding fails to recover for instance one atom, $I^t = I_{i\leftrightarrow j}$. Then we still have $a = Q(\Phi_{I^t})(x_i\phi_i + r) \approx x_i\phi_i + r$ and $|\langle \phi_i, a\rangle|^2 \gtrsim \theta_K^2 \|a\|_2$. If thresholding constantly misses one atom in the support, for instance because the current dictionary estimate is quite coherent, $\mu \gg 1/\sqrt{d}$, or not yet very accurate, this will lead to an increase $S_e = S + 1$. However, as we have discussed above, while increasing the sparsity level increases the chances for full recovery by thresholding, it also increases the atom estimation error, which decreases the chances for full recovery. Depending on which effect dominates, this could lead to a vicious circle of increasing the sparsity level, which decreases the accuracy leading to more failure of thresholding and increasing the sparsity level. In order to avoid this risk, we should take into account that thresholding might fail to recover the full support and be able to identify such failure. Further, we should be prepared to also decrease the sparsity level.

The key to these three goals is to also look at the coefficients of the signal approximation. Assume that we are given the correct sparsity level $S_e = S$ but recovered $I^t = I_{i\leftrightarrow j}$. Defining $I_{i\rightarrow} = I \setminus \{i\}$, the corresponding coefficients $\tilde{x}_{I^t}$ have the shape,

$$\begin{aligned} \tilde{x}_{I^t} = \Phi_{I^t}^\dagger(\Phi_I x_I + r) &= \Phi_{I^t}^\dagger(\Phi_{I_{i\rightarrow}} x_{I_{i\rightarrow}} + \phi_i x_i + r) \\ &= (x_{I_{i\rightarrow}}, 0) + (\Phi_{I^t}^\star \Phi_{I^t})^{-1}\Phi_{I^t}^\star(\phi_i x_i + r), \end{aligned} \tag{4.6}$$

meaning $|\tilde{x}_{I^t}(j)|^2 \leq (\mu^2|x_i|^2 + |\langle\phi_j, r\rangle|^2)/(1-\mu S)^2$ or even $|\tilde{x}_{I^t}(j)|^2 \lesssim \mu^2|x_i|^2 + |\langle\phi_j, r\rangle|^2$. Since the residual is again approximately $a \approx \phi_i x_i + r$, this means that for incoherent dictionaries, the coefficient of the wrongly chosen atom is likely to be below the threshold $\theta_K^2\|a\|_2$, while the one of the missing atom will be above the threshold, so we are likely to keep the sparsity level the same. Similarly, if we overestimate the sparsity level $S_e = S + 1$ and recover an extra atom $I^t = I_{\leftarrow j} := I \cup \{j\}$, we have $a = Q(\Phi_{I^t})r \approx r$ while the coefficient of the extra atom will be of size $|\tilde{x}_{I^t}(j)|^2 \approx |\langle\phi_j, r\rangle|^2 < \theta_K^2\|a\|_2^2$. All in all our estimates suggest that we get a more stable estimate of the sparsity level by averaging the number of coefficients $\tilde{x}_{I^t} = \Phi_{I^t}^\dagger y$ and residual inner products $(\langle\phi_i, a\rangle)_{i\notin I^t}$ that have squared value larger than θ_K^2 times the residual energy. However, the last detail we need to include in our considerations is the reason for thresholding failing to recover the full support given the correct sparsity level in first place. Assume for instance, that the signal does not contain noise, $y = \Phi_I x_I$, but that the sparse coefficients vary quite a lot in size. In Subsection 3.2.1 we have seen that in case of i.i.d. random coefficient signs, $\mathbb{P}(\mathrm{sign}(x_i) = 1) = 1/2$, the inner products of the atoms inside resp. outside the support concentrate around,

$$\begin{aligned} &i \in I \qquad |\langle\phi_i, \Phi_I x_I\rangle| \approx |x_i| \pm \left(\textstyle\sum_{k\neq i} x_k^2 |\langle\phi_i, \phi_k\rangle|^2\right)^{1/2} \approx |x_i| \pm \mu\|y\|_2 \\ &i \notin I \qquad |\langle\phi_i, \Phi_I x_I\rangle| \approx \left(\textstyle\sum_k x_k^2 |\langle\phi_i, \phi_k\rangle|^2\right)^{1/2} \approx \mu\|y\|_2. \end{aligned}$$

This means that thresholding will only recover the atoms corresponding to the S_r-largest coefficients for $S_r < S$, that is, $I_r = \{i \in I : |x_i| \gtrsim \mu\|y\|_2\}$. The good news is that these will capture most of the signal energy, $\|P(\Phi_{I^t})y\|_2^2 \approx \|\Phi_{I_r} x_{I_r}\|_2^2 \approx \|y\|_2^2$, meaning that in some sense the signal is only S_r sparse. It also means that for $\mu^2 \approx 1/d$, we can estimate the *recoverable* sparsity level of a given signal as the number of squared coefficients/residual inner products that are larger than

$$\frac{1}{d}\|P(\Phi_{I^t})y\|_2^2 + \frac{2\log(4K)}{d}\|Q(\Phi_{I^t})y\|_2^2. \tag{4.7}$$

If S_n is the estimated recoverable sparsity level of signal y_n, a good estimate of the overall sparsity level S will be the rounded average sparsity level $\bar{S} = \lfloor \frac{1}{N}\sum_n S_n \rceil$. The corresponding update rule then is to increase S_e by one if $\bar{S} > S_e$, keep it the same if $\bar{S} = S_e$ and decrease it by one if $\bar{S} < S_e$, formally

$$S_e^{new} = S_e + \operatorname{sign}(\bar{S} - S_e). \tag{4.8}$$

In the next subsection, we address the big question of how to adaptively select the dictionary size K.

4.3.2 Adapting the dictionary size

The choice of the dictionary size K might be motivated by a budget, such as being able to store K atoms and S values per signal, or application specific, that is, the expected number of sources in sparse source separation. In applications such as image restoration K (like S) is either chosen ad hoc or experimentally with an eye towards computational complexity, and one will usually find $d \leq K \leq 4d$, and $S = \sqrt{d}$. If algorithms include some sort of adaptivity of the dictionary size, this is usually in the form of not updating unused atoms, a rare occurrence in noisy situations, and deleting them at the end. Also this strategy can only help if K was chosen too large but not if it was chosen too small.

Underestimating the size of a dictionary obviously prevents recovery of the generating dictionary. For instance, if we provide $K-1$ instead of K, the best we can hope for is a dictionary containing $K-2$ generating atoms and a $1:1$ combination of the two missing atoms. The good news is that if we are using a replacement strategy, one of the candidates will encode the $1:1$ complement, similar to the situation discussed in the last section where we are given the correct dictionary size but had a double atom. Overestimating the dictionary size does not prevent recovering the dictionary per se, but can decrease the recovery precision, meaning that a bigger dictionary might not actually provide a smaller approximation error. To get an intuition what happens in this case, assume that we are given a budget of $K+1$ instead of K atoms and the true sparsity level S. The most useful way to spend the extra budget is to add a $1:1$ combination of two atoms, which frequently occur together, meaning $\phi_0 \propto \phi_i + h\phi_j$ for $h = \operatorname{sign}(\langle \phi_i, \phi_j \rangle)$. The advantage of the augmented dictionary $\Psi = (\phi_0, \Phi)$ is that some signals are now $S-1$ sparse. The disadvantage is that Ψ is less stable since the extra atom ϕ_0 will prevent ϕ_i or ϕ_j to be selected by thresholding whenever they

are contained in the support in a $1:h$ ratio. This disturbs the averaging process and reduces the final accuracy of both ϕ_i and ϕ_j.
The good news is that the extra atom ϕ_0 is actually quite coherent with the dictionary, $\left|\langle\phi_0, \phi_{i(j)}\rangle\right| \geq 1/\sqrt{2}$, so if we have activated a replacement threshold of $\mu_{\max} \leq 1/\sqrt{2}$, the atom ϕ_0 will be soon replaced, necessarily with another useless atom.
This suggests as strategy for adaptively choosing the dictionary size to decouple our replacement scheme into pruning and adding, which allows to both increase and decrease the dictionary size. We will first have a closer look at pruning.

Pruning atoms

From the replacement strategy we can derive two easy rules for pruning: 1) if two atoms are too coherent, delete the less often used one or merge them, 2) if an atom is not used, delete it. Unfortunately, the second rule is too naive for real world signals, containing among other imperfections noise, which means also purely random atoms are likely to be used at least once by mistake. To see how we need to refine the second rule assume again that our sparse signals are affected by Gaussian noise (of a known level), that is, $y = \Phi_I x_I + r$ with $\mathbb{E}(\|r\|_2^2) = \rho^2$ and that our current dictionary estimate has the form $\Psi = (\phi_0, \Phi)$, where ϕ_0 is some vector with admissible coherence to Φ. Whenever ϕ_0 is selected this means that thresholding has failed. From the last subsection we also know that we have a good chance of identifying the failure of thresholding by looking at the coefficients $\Phi_{I^t}^\dagger(\Phi_I x_I + r)$. The squared coefficient corresponding to the incorrectly chosen atom ϕ_0 is likely to be smaller than $\lesssim \|\Phi_I x_I\|_2^2/d + |\langle\phi_0, r\rangle|^2$ while the squared coefficient of a correctly chosen atom $i \in I \cap I^t$ will be larger than $|x_i|^2 + |\langle\psi_i, r\rangle|^2 \gtrsim \|\Phi_I x_I\|_2^2/S + |\langle\phi_i, r\rangle|^2$, at least half of the time. The size of the inner product of any atom with Gaussian noise can be estimated as

$$\mathbb{P}\left(|\langle\phi_k, r\rangle| > \tau\|r\|_2\right) \leq 2\exp\left(-\frac{d\tau^2}{2}\right). \tag{4.9}$$

Taking again $\|P(\Phi_{I^t})y\|_2$ as estimate for $\|\Phi_I x_I\|_2$ and $\|a\|_2 = \|Q(\Phi_{I^t})y\|_2$ as estimate for $\|r\|_2$, we can define the refined value function $\tilde{v}(k)$ as the number of times an atom ϕ_k has been selected and the corresponding coefficient has squared value larger than $\|P(\Phi_{I^t})y\|_2^2/d + \tau^2\|a_n\|_2^2$. Based on the bound above we can then estimate that for N noisy signals the value function of the unnecessary or random atom ϕ_0 is bounded by $\tilde{v}(0) \lesssim 2N\exp\left(-\frac{d\tau^2}{2}\right) =: M$, leading to a natural criterion for deleting unused atoms. Setting for instance $\tau = \theta_K = \sqrt{2\log(4K)/d}$, we get $M = N/(2d)$. Alternatively, we can say that in order to accurately estimate an atom, we need M reliable observations and accordingly set the threshold to $\tau = \sqrt{2\log(2N/M)/d}$.
The advantage of a relatively high threshold $\tau \approx \sqrt{2\log(4K)/d}$ is that in low noise scenarios, we can also find atoms that are rarely used. The disadvantage is that for high τ the quantities $\tilde{v}(\cdot)$ we have to estimate are relatively small and therefore susceptible

to random fluctuations. In other words, the number of training signals N needs to be large enough to have sufficient concentration such that for unnecessary atoms the value function $\tilde{v}(\cdot)$ is actually smaller than M. Another consideration is that at the beginning, when the dictionary estimate is not yet very accurate, also the approximate versions of frequently used atoms will not be above the threshold often enough. This risk is further increased if we also have to estimate the sparsity level. If S_e is still small compared to the true level S we will overestimate the noise, and even perfectly balanced coefficients $1/\sqrt{S}$ will not yet be above the threshold. Therefore, pruning of the dictionary should only start after an embargo period of several iterations to get a good estimate of the sparsity level and most dictionary atoms.
In the replacement section we have also seen that if a double atom is replaced by the 1:1 complement $\phi_i - \phi_j$ of a 1:1 atom $\phi_i + \phi_j$, it will take a few iterations for the pair $(\phi_i \pm \phi_j)$ to rotate into the correct configuration (ϕ_i, ϕ_j), where they are recovered most of the time. In the case of decoupled pruning and adding, we run the risk of deleting a missing atom or a $1:1$ complement one iteration after adding it simply because it has not been used often enough. Therefore, every freshly added atom should not be checked for its usefulness until after a similar embargo period of several iterations, which leads right to the next question when to add an atom.

Adding atoms

To see when we should add a candidate atom to the dictionary, we have a look back at the derivation of the replacement strategy. There we have seen that the residuals are likely to be either 1-sparse in the missing atoms (or 1:1 complements of the atoms doing the job of two generating atoms), meaning, $a \approx |x_i|/2(\phi_i - \phi_j)$ or in a more realistic situation $a \approx |x_i|/2(\phi_i - \phi_j) + r$, or zero, which again in the case of noise means $a \approx r$. To identify a good candidate atom we observe again that if the residual consists only of (Gaussian) noise, we have for any vector/atom γ_k

$$\mathbb{P}\left(|\langle \gamma_k, r\rangle| > \tau_\Gamma \|r\|_2\right) \leq 2\exp\left(-\frac{d\tau_\Gamma^2}{2}\right). \tag{4.10}$$

If on the other hand the residual consists of a missing complement, the corresponding candidate $\gamma_\ell \approx (\phi_i - \phi_j)/\sqrt{2}$ should have $|\langle a, \gamma_\ell\rangle| \approx |x_i|/\sqrt{2} \gtrsim \tau_\Gamma \|a\|_2$. This means that we can use a similar strategy as for the dictionary atoms to distinguish between useful and useless candidates. In the last candidate iteration, using N_Γ residuals, we count for each candidate atom γ_k how often it is selected and satisfies $|\langle \gamma_k, a\rangle| > \tau_\Gamma \|a\|_2$. Following the dictionary update and pruning, we then add all candidates to the dictionary whose value function is higher than $M_\Gamma = 2N_\Gamma \exp\left(-\frac{d\tau_\Gamma^2}{2}\right)$ and which are incoherent enough to atoms already in the dictionary.

Now that we have covered all aspects that are necessary for making ITKrM adaptive, in the next chapter we will test its performance in a real-world application. More specifically, we will use this adaptive version of ITKrM for the reconstruction of cardiac cine MR images from highly undersampled data. For the ones interested in numerical experiments testing its performance on synthetic data and small image data we refer to the original paper [49].

Chapter 5

Adaptive Dictionary Learning for MR Image Reconstruction

In this chapter we present an application of the adaptive version of the ITKrM algorithm (aITKrM) to the reconstruction of cardiac MR images. In particular, for our results we use aITKrM for learning the dictionary and introduce an adaptive version of OMP (aOMP) which we use for sparse coding. We conduct several experiments to show the competitiveness and advantages of these adaptive algorithms compared to non-adaptive methods. The results obtained also show the difficulty of choosing the correct sparsity level S and dictionary size K, and thus the importance of their adaptive choice. The material presented in this chapter has been published in [48].

5.1 Introduction to MRI

Magnetic Resonance Imaging (MRI) has become nowadays an indispensable imaging modality which is widely used in daily clinical routine to image the interior of a patient. For example, cardiac cine MRI can be used for the assessment of the cardiac function. For that, a slice of the patient's heart is scanned over multiple cardiac cycles and a sequence of 2D (2 dimensional) images showing the heart movement can be obtained. However, a major issue of MRI is the slow data-acquisition process due to physical limits imposed by the scanner. In particular, typical cardiac MR-scans are performed during a breathhold to avoid respiratory motion artefacts. Therefore the breathhold duration limits the spatial and temporal resolution of MR-scans, which represents a problem for ill patients with limited breathhold capabilities. The data-acquisition in MRI takes place in the so-called k-space, i.e. the Fourier space. Since the acquisition is often slow, undersampling in k-space is used to shorten scan times. This leads to undersampling artefacts due to the violation of the Nyquist sampling limit. Parallel imaging and regularised iterative reconstruction methods have been proposed to minimise undersampling artefacts, e.g. [77]. Regularisation approaches using transforms learned from data, meaning, dictionary learning and sparse coding (the sparse approximation of the data) have been considered in the past [13, 52, 39, 72, 70, 9, 74, 7, 73, 63]. In dictionary learning-based regularisation, the model assumption is patch-wise sparsity and therefore, the idea is to patch-wise impose the regularisation on the image to be reconstructed.

The rationale behind the regularisation based on learned dictionaries is that patches of an image have an inherently low-dimensional representation and therefore, noise-like components of an image can be removed by sparsely approximating the image-patches with respect to a learned dictionary, see e.g. [70]. The regularisation of the solution is achieved by the fact that, given the incoherent undersampling scheme applied in k-space, the artefacts resulting from the direct reconstruction of an image are high-dimensional and thus suppressed by the low-dimensional representation, which suffices to capture the important features.

In [13], for example, a pre-trained dictionary is used to regularise the images. In [71], regularisation using a pre-trained dictionary across multichannels is used for calibration-free parallel MR imaging. Further, approaches in which the dictionary is learned from the current image estimate during the reconstruction have been proposed [52, 39] and successfully applied to cine MR image reconstruction [9, 74]. However, regardless of the excellent image quality which can be achieved by the latter mentioned methods, there still remain a few issues. First, the sparsity level S used for dictionary learning and sparse coding as well as the number of atoms in the dictionary K need to be chosen a-priori and are typically determined by repeating the experiments for different choices of S and K. However, the parameters are clearly data-dependent and there is no guaranty on the achievable performance of the reconstruction algorithms on different datasets. Second, performing an S-sparse approximation of all image patches

is computationally quite expensive, especially when S is chosen relatively high. These two issues make the method prohibitive for the application in the clinical routine where standardised reconstruction protocols have to be used.
To overcome the problem given by the computational complexity of the dictionary learning- and sparse coding-stage as well as the need for choosing S and K, we will use adaptive versions of dictionary learning and sparse approximation algorithms. In particular, for learning the dictionary we use the adaptive version of the ITKrM algorithm which we discussed in Chaper 4, and show its competitiveness against the well-known K-SVD algorithm. For sparse coding we introduce an adaptive version of OMP which is based on the selection of the atoms using thresholding, similar to [16] and [20]. However, while [16] and [20] require the careful tuning of a hyper-parameter, our choice of the threshold is inspired by the considerations in Chapter 4 and hence, selected based on the dictionary size K.

This chapter is structured as follows. In Section 5.2 we formulate the reconstruction problem using dictionary learning and introduce an adaptive version of OMP. In Section 5.3 we describe different experiments, where the obtained results are presented in Section 5.4 and discussed in Section 5.5.

5.2 Problem Formulation and Dictionary Learning-based Regularisation Approaches

Mathematically, the process of undersampling can be formulated as applying a binary mask $\mathbf{S}_U$ to the measured Fourier data. Let $\mathbf{y} \in \mathbb{C}^{N_F}$ denote the vector representation of the unknown cine MR image with $N_F = N_x \cdot N_y \cdot N_t$, where $N_x \times N_y$ is the shape of a single 2D image and N_t corresponds to the number of cardiac phases. Let $\mathbf{F}$ denote the encoding operator and $U \subset J = \{1, \dots, N_F\}$ the set of Fourier coefficients which are needed to properly reconstruct the image $\mathbf{y}$. The inverse problem one aims to solve is of the form

$$\tilde{\mathbf{y}}_U = \mathbf{F}_U \mathbf{y} + \eta, \tag{5.1}$$

where $\mathbf{F}_U := \mathbf{S}_U \circ \mathbf{F}$ and η denotes random noise. Images directly reconstructed from undersampled k-space by applying the adjoint operator $\mathbf{F}_U^{\mathsf{H}}$ contain severe artefacts which limit the diagnostic quality. Since by discarding non-measured data the problem (5.1) becomes underdetermined, i.e. there is an infinite number of solutions. In Parallel Imaging, where multiple receiver coils are used, the system can be overdetermined but the reconstruction problem becomes poorly conditioned. Therefore, in order to constrain the space of solutions of interest, regularisation techniques must be used. When dictionary learning and sparse coding are used as a regularisation method, possible formulations of the image reconstruction problem are the ones of the joint minimisation problems, for a fixed dictionary Ψ

$$\min_{\mathbf{y},\{x_j\}_j} \|\mathbf{F}_U\mathbf{y} - \tilde{\mathbf{y}}_U\|_2^2 + \frac{\lambda}{2}\sum_j \left(\|\mathbf{R}_j\mathbf{y} - \Psi x_j\|_2^2 + \|x_j\|_0\right), \tag{P1}$$

see e.g. [13], or with minimisation also over the dictionary

$$\min_{\mathbf{y},\Psi,\{x_j\}_j} \|\mathbf{F}_U\mathbf{y} - \tilde{\mathbf{y}}_U\|_2^2 + \frac{\lambda}{2}\sum_j \left(\|\mathbf{R}_j\mathbf{y} - \Psi x_j\|_2^2 + \|x_j\|_0\right), \tag{P2}$$

see e.g. [9] and [74]. Here, $\mathbf{y}$ denotes the unknown solution, $\tilde{\mathbf{y}}_U$ the measured undersampled acquired k-space data, λ a regularisation parameter, and $\|x_j\|_0$ counts the number of non-zero coefficients in x_j. The operator $\mathbf{R}_j$ extracts the j-th 3D patch from the image $\mathbf{y}$, Ψ denotes the dictionary and x_j the sparse coefficient vector of the patch $\mathbf{R}_j\mathbf{y}$ with respect to Ψ. The difference between (P1) and (P2) is that in (P1), one assumes to have a pre-trained dictionary Ψ, while in (P2), the dictionary Ψ is learned during the reconstruction based on the current image estimates. Note that in [74] and [9], a Total Variation (TV) term is further added to the minimisation problem (P2). However, since we want to focus on the dictionary learning component of the reconstruction, we neglect the additional TV-regularisation term. Problems (P1) and (P2) can be solved by the alternating direction method of multipliers (ADMM) which alternates between the minimisation with respect to $\mathbf{y}$, the dictionary Ψ and the set of vectors $\{x_j\}_j$. Usually, the starting point for the iterative reconstruction algorithm is given by the direct reconstruction from the measured data, that is $\mathbf{y}_U = \mathbf{F}_U^{\mathsf{H}}\tilde{\mathbf{y}}_U$.

Dictionary and Sparse Code Update

Assuming a fixed $\mathbf{y}$, the minimisation of (P1) and (P2) is achieved by solving the problems

$$\min_{\{x_j\}_j} \sum_j \left(\|\mathbf{R}_j\mathbf{y} - \Psi x_j\|_2^2 + \|x_j\|_0\right) \tag{5.2}$$

and

$$\min_{\Psi,\{x_j\}_j} \sum_j \left(\|\mathbf{R}_j\mathbf{y} - \Psi x_j\|_2^2 + \|x_j\|_0\right), \tag{5.3}$$

respectively. Problem (5.2) is solved (approximately) with any sparse approximation algorithm, while (5.3) is typically solved using first dictionary learning to learn Ψ and then sparse coding. Note that, if the dictionary is learned using an alternating minimisation algorithm, which alternates between dictionary learning to obtain Ψ and sparse coding to obtain the vectors x_j, the sparse coding after dictionary learning could be skipped. However, usually we have to use different sparsity levels in the dictionary learning- and sparse coding-stage or different sparse approximation algorithms in and after dictionary learning are used.

Reconstruction Update

Assuming a fixed dictionary Ψ and a fixed set of sparse coefficient vectors $\{x_j\}_j$, one can easily see that minimising (P1) or (P2) with respect to $\mathbf{y}$ is equivalent to solving the system of linear equations

$$\mathbf{H}\mathbf{y} = \mathbf{b}, \tag{5.4}$$

where the operator $\mathbf{H}$ is given by

$$\mathbf{H} = \mathbf{F}_U^{\mathsf{H}}\mathbf{F}_U + \lambda \sum_j \mathbf{R}_j^{\mathsf{T}}\mathbf{R}_j, \tag{5.5}$$

and the right-hand-side $\mathbf{b}$ is given by a linear combination of the initial reconstruction $\mathbf{y}_U$ and an image which corresponds to a properly averaged combination of all patches Ψx_j, i.e.

$$\mathbf{b} = \mathbf{F}_U^{\mathsf{H}}\tilde{\mathbf{y}}_U + \lambda \sum_j \mathbf{R}_j^{\mathsf{T}}\Psi x_j. \tag{5.6}$$

Since in general, the inversion of the operator $\mathbf{H}$ is computationally not feasible, problem (5.4) is solved using an iterative method. Given that $\mathbf{H}$ is symmetric, a common choice for the solver is the pre-conditioned conjugate gradient (PCG) method [28].

5.2.1 Adaptive dictionary learning and sparse coding algorithms

In applications such as image restoration, the sparsity level S and the dictionary size K are typically chosen empirically or experimentally. As already mentioned in the previous chapter, for d-dimensional signals, typical values are $d \leq K \leq 4d$ and $S = \sqrt{d}$, but depending on the situation they can highly vary and, as we will show later, they might have a significant impact on the reconstruction quality as well as the required computational time.

To circumvent this issue and in order to investigate its performance in practical applications, we use the adaptive version of ITKrM (aITKrM). Inspired by some of the ideas which we used for adapting the sparsity level S in Chapter 4, we now further introduce adaptivity in the sparse coding stage after dictionary learning. In particular, in the following, we present an adaptive version of OMP where not only the sparsity level S is chosen adaptively but which will also turn out to significantly accelerate the sparse coding procedure. As we will demonstrate later, the sparsity level of an image can vary from position to position, meaning, depending on the texture of each image patch, we have higher or lower S, hence, suggesting to introduce an adaptive choice of S also in the sparse coding step.

Adaptive OMP

In order to incorporate adaptivity into OMP, we replace the condition of stopping after adding at most S atoms by a bound for the maximal inner product between any atom and the current residual. More precisely, in each iteration, we check if there exists an atom ψ_k for which the absolute value of the residual inner product $|\langle \psi_k, a_n\rangle|$ is larger than some threshold times the norm of the residual. The index corresponding to the atom yielding the largest inner product is then selected. Projecting the signal onto the span of already selected atoms and calculating the new residual, this procedure is repeated until the stopping condition is met. A suitable threshold is obtained using concentration of measure. More precisely, we want aOMP to stop if the residual consists only of noise. For that, assume our current residual is of the form $a_n = r$, where r denotes a Gaussian noise vector, and for the current support $|I_n| = S$. The expected number of remaining atoms for which the residual inner product is larger than $\tau\|r\|_2$ can be calculated as

$$\begin{aligned}\mathbb{E}\big(\sharp\{k \notin I_n : |\langle \psi_k, r\rangle| > \tau\|r\|_2\}\big) &= \sum_{k\notin I_n} \mathbb{P}\big(|\langle \psi_k, r\rangle| > \tau\|r\|_2\big)\\ &< 2(K-S)\exp\left(-\frac{d\tau^2}{2}\right). \qquad (5.7)\end{aligned}$$

Setting $\tau = \sqrt{2\log(4K)/d}$, the expectation above is smaller than $\frac{1}{2}$. This means, if the residuals consist only of noise, using $\tau = \sqrt{2\log(4K)/d}$ within the stopping condition of aOMP, on average half an atom per signal has a residual inner product larger than $\tau\|r\|_2$. This expectation can be further decreased for an even higher choice of the threshold τ. Inequality (5.7) is the main ingredient of the algorithm as it provides a threshold τ that prevents aOMP from overfitting the considered patches and removes noise.

To further accelerate aOMP, we introduce a preliminary step where we select the *strongest* part of the support using a slightly higher threshold than τ. In particular, before always adding the next best fitting atom (one at a time) we will choose part of the support (several atoms at a time) by thresholding with $\tau_1 = \sqrt{2\log(8K)/d}$. This partial support is subsequently refined/expanded by proceeding aOMP until one of the stopping conditions is met. A summary of the proposed algorithm can be found in Algorithm 5.2.1.

Note that we suggest to use OMP for the sparse coding stage of the iterative reconstruction however, not to replace thresholding by OMP within aITKrM. This choice is motivated by two reasons. First, thresholding is computational much cheaper than OMP and hence, suitable for accelerating the regularisation stage of the iterative reconstruction. Second, although OMP is known to yield better results than thresholding for sparse approximation, it is unstable under perturbations. More precisely,

Algorithm 5.2.1: Adaptive Orthogonal Matching Pursuit (aOMP)

```
Input: $\Psi, Y$ ;                                  // dictionary, signals
Initialise: $X = 0$ ;                               // $d \times N$ matrix
$\tau_1 = \sqrt{2\log(8K)/d}$ ;                     // thresholds
$\tau_2 = \sqrt{2\log(4K)/d}$
foreach $n$ do
    $I_n^t = \arg\text{where}\big(|\langle\psi_k, y_n\rangle| > \tau_1 \cdot \|y_n\|_2\big)$
    $a_n = y_n - P(\Psi_{I_n^t})y_n$
    while $\max_k |\langle\psi_k, a_n\rangle| > \tau_2 \cdot \|a_n\|_2$ do
        $I_n^t = I_n^t \cup \arg\max_k |\langle\psi_k, a_n\rangle|$
        $a_n = y_n - P(\Psi_{I_n^t})y_n$
    end
    $X[I_n^t, n] = \Psi_{I_n^t}^\dagger y_n$
end
Output: $X$ ;                                       // sparse coefficient matrix
```

using an appropriate dictionary Ψ for the sparse approximation of a class of signals, OMP is known to yield much better results than simple thresholding. However, in the presence of perturbations of the dictionary, which is the case during learning the dictionary, OMP performs worse. In order to verify this claim, in Chapter 6 we investigate the performance of OMP in case where the input dictionary is only a perturbed version of the generating dictionary.

5.3 In-Vivo Experiments

Here, we describe the experiments which we conducted in order to study the behaviour of aITKrM and aOMP when they are used for the reconstruction of 2D cine MR images from undersampled k-space data.
In particular, in order to get an assessment of the quality of the obtained reconstructions for various combinations of dictionary learning and sparse approximation algorithms and to highlight some aspects of the adaptive dictionary learning and sparse approximation algorithms, we performed the following experiments.

1. *Adaptive vs. non-adaptive dictionary learning and sparse coding:* Here, we quantitatively compared the performance of the reconstruction algorithms used to solve problems (P1) and (P2) using three different combinations of dictionary learning and sparse approximation algorithms: K-SVD + OMP, ITKrM + OMP

and aITKrM + aOMP. For these experiments, images obtained by kt-SENSE [67] were used as ground-truth images. From these images, the k-space data was retrospectively generated and corrupted by Gaussian noise in order to simulate an acceleration factor of 9. We repeated the experiments for different choices of the sparsity level S. More precisely, to demonstrate the impact of the choice of potentially too low/too high S, we used $S = 4$, $S = 8$ and $S = 16$ for the non-adaptive dictionary learning and sparse approximation algorithms.

2. *Convergence behaviour:* We investigated the convergence behaviour of the different combinations of dictionary learning and sparse coding methods by tracking the average of the chosen image measures during the reconstruction.

3. *Computational time:* We compared the different combinations K-SVD + OMP / ITKrM + OMP / aITKrM + aOMP in terms of computational time.

4. *Sensitivity with respect to $\mu_{\max}$ and M:* Since for aITKrM the maximal allowed coherence of the dictionary $\mu_{\max}$ and the minimal number of observations M have to be chosen, we have compared the obtained results for different choices of M and $\mu_{\max}$ to demonstrate the stability with respect to them.

5. *Experiments using real k-space data:* Here, we reconstructed images from the k-space data obtained from the scanner with the three different combinations of dictionary learning and sparse coding.

For all experiments, we used the publicly available `Python`-implementations of K-SVD and OMP in the `scikit-learn` library [51] which are based on an efficient implementation of K-SVD using batch OMP [54]. The forward and the adjoint operators $\mathbf{F}_U$ and $\mathbf{F}_U^{\mathsf{H}}$ were implemented using the libraries `ODL` [1] and `PyNUFFT` [38]. The PCG method used to solve system (5.4) was provided by `ODL`. Our `Python`-implementations of ITKrM, aITKrM and aOMP as well as of the forward and adjoint operators $\mathbf{F}_U$ and $\mathbf{F}_U^{\mathsf{H}}$ using the library `Torch KB-NUFFT` [46], [45], are available at `https://github.com/koflera/AdaptiveDLMRI`.

5.3.1 Dataset

Our dataset consisted of $n = 15$ 2D cine MR image series from patients as well as healthy volunteers and represents a subset of particularly interesting cases selected from [31]. Further, 10 different images were used to pre-train dictionaries for solving (P1). The images were obtained using a bSSFP sequence on a 1.5 T MR scanner (Achieva, Philips Healthcare, Best, The Netherlands) within a single breathhold of 10 s (TR/TE = 3.0/1.5 ms, FA 60°). The images have a shape of $N_x \times N_y \times N_t = 320 \times 320 \times 30$, where $N_x \times N_y$ is the number of in-plane pixels and N_t is the number of cardiac phases which were acquired during the scan. The in-plane resolution of the images is 2 mm and the slice thickness is 8 mm. The acquired k-space data corresponds to the Fourier-data sampled along $N_\theta = 3400$ radial trajectories which were chosen according to [76]. From these images, we retrospectively generated the undersampled k-space data $\tilde{\mathbf{y}}_U$ by solely using $N_\theta = 1130$ radial spokes. Using only $N_\theta = 1130$ spokes corresponds to an undersampling factor of approximately ~ 9 and reduces the scan time to approximately 3 seconds. Further, the k-space data was corrupted by a normally distributed random noise vector η with a standard deviation of 0.05.

5.3.2 Experiment set-up

The patch-size used for all experiments was $4 \times 4 \times 4$, which we chose according to other published methods, e.g. [9, 74]. Our signals, which correspond to the vectorised patches extracted from the images, therefore have dimension $d = 64$. As in [9], we approximated the real and imaginary part of the complex-valued images separately but used the same real-valued dictionary Ψ. For the non-adaptive combinations of dictionary learning and sparse approximation algorithms, we fixed the number of atoms of the dictionary Ψ to be $K = 128$. Note that the empirical choice of K is typically in the range $d \leq K \leq 4d$ while using a sparsity level of $S = \sqrt{d}$, which, for a fixed size of patches $4 \times 4 \times 4$, results in $64 \leq K \leq 256$ and $S = 8$. In fact, this choice is well-established in the literature. For example, in [74], the parameters are empirically set to $K = 256$ and $S = 15$. In [9], the number of atoms is set even higher, varying from $K = 300$ to $K = 600$, dependent on the experiments. However, due to the fact that our forward model is given by a radial encoding operator using multiple coils, the artefacts contained in the initial reconstruction $\mathbf{y}_U$ which is obtained using the non-uniform fast Fourier transform (NUFFT) are inherently different from the ones obtained by a zero-filled reconstruction as in [9] or [74]. Since the artefacts can be expected to have a more high-frequency type of texture, we decided to only use $K = 128$ for varying S. As already mentioned, the experiments were repeated for a relatively low choice of $S = 4$, a typical choice $S = \sqrt{d} = 8$ and a relatively high choice of $S = 16$. Further, for $S = 8$, the number of atoms K is also varied in other experiments. The sparsity level S was chosen to be the same for both dictionary learning and sparse coding. Note that, this choice is known to be sub-optimal, e.g. for the combination ITKrM + OMP, where usually a smaller S for learning the dictionary and a larger S for sparse coding is required. However, for the ease of comparison they were chosen to be the same. In any case our point is that a global choice of the sparsity level can never be optimal. Since the k-space data $\tilde{\mathbf{y}}_U$ was contaminated by random noise, the regularisation parameter λ was set to $\lambda = 1$ in order to achieve a relatively strong contribution of the regularisation imposed by dictionary learning and sparse coding and therefore being able to highlight the impact of the different dictionary learning and sparse approximation algorithms. The number of PCG iterations used to update the reconstruction by solving (5.4) and the number of overall iterations for

ADMM were set to $n_{\mathrm{PCG}} = 4$ and $T = 12$, respectively.
For solving (P1), the dictionaries were pre-trained on patches extracted from the images of 10 different subjects. The dictionaries were initialised by $K = 128$ randomly selected patches and learned by randomly extracting 150 000 patches of the real and imaginary part of the images at each dictionary learning iteration. The maximal number of iterations for the respective dictionary learning algorithm was set to $n_{\mathrm{DL}} = 200$. For the combination aITKrM + aOMP, we used $\mu_{\max} = 0.7$ and for the number of minimal observations we used $M = d$ with corresponding coefficient threshold $\tau = \sqrt{2\log(2N/M)/d}$. During every iteration of aITKrM we learned $L = \lfloor \log d \rceil = 4$ replacement candidates using $m = \lfloor \log d \rceil = 4$ iterations each with $N_\Gamma = \lfloor N/m \rfloor$ signals. Also for the number of protected runs for newly added atoms we chose $m = \lfloor \log d \rceil = 4$. Promising replacement candidates were added to the dictionary after every iteration starting with the m-th iteration. In the last $3m$ iterations no more atoms were added. Coherent atoms were merged after every iteration and unused atoms were pruned after every iteration starting with iteration $2m$. The resulting size of the dictionary learned with aITKrM was $K = 151$.
For solving (P2), the dictionaries were learned by randomly extracting $N = 10\,000$ patches of the real and imaginary part of the current image estimate $\mathbf{y}_k$ for each dictionary learning iteration. The maximal number of iterations of the respective dictionary learning algorithm within one ADMM iteration was set to $n_{\mathrm{DL}} = 20$. The dictionaries were initialised as for solving (P1) and continuously updated during the reconstruction. For each subsequent ADMM iteration, the dictionary Ψ was initialised with the one learned during the previous ADMM iteration. The set-up for the adaptive part was the same as for solving (P1). For sparse approximation we used strides of 2 in N_x-, N_y- and N_t-direction, which reduces the number of patches to be sparsely approximated by a factor of 8. Note that we did not learn the constant atom since the patches were centered before learning the dictionaries.

5.3.3 Quantitative measures

For evaluating the performance of the different reconstruction algorithms, we report the peak signal-to-noise ratio (PSNR) and the normalised root mean squared error (NRMSE) as error-based image metrics and the structural similarity index measure [75] (SSIM) as similarity-based image metric. The hyper-parameters needed by SSIM are the ones published in the respective work. In order to focus on the regions of the images with diagnostic content, the metrics were calculated on the images which were previously cropped to $N_x' \times N_y' = 220 \times 220$.

5.4 Results

In this section we present the results which we obtained from the various experiments described above.

5.4.1 Reconstruction results

Here, we reconstructed all 15 cine MR image series using the different combinations of dictionary learning and sparse approximation algorithms. Figure 5.1 shows an example of images reconstructed with the three different combinations of dictionary learning and sparse approximation algorithms. For K-SVD and ITKrM, the combination of S and K is the one which led to the best quantitative results for the respective methods. Concretely, when the dictionary is pre-trained the best choice for K-SVD was $S = 8$ and $K = 128$, and for ITKrM $S = 8$ and $K = 64$. As can be seen from the point-wise error images, all non-adaptive and the adaptive dictionary learning and sparse coding combinations led to visually comparable results. Table 5.1 lists the average PSNR, NRMSE and SSIM for the different reconstructions. When the dictionary Ψ is learned during the reconstruction, we see that for both non-adaptive combinations K-SVD + OMP and ITKrM + OMP, setting $S = 16$ yielded the worst results compared to $S = 8$ and $S = 4$. In particular, the gap between them was larger for larger S, which can be attributed to issues during the dictionary learning and is a well known issue of ITKrM for overestimated sparsity levels [61]. The adaptive combination aITKrM + aOMP achieved similar reconstruction quality as K-SVD + OMP with the best reported choices of the sparsity level S and dictionary size K by further slightly improving SSIM.

The second part of Table 5.1 lists the results obtained by using a pre-trained dictionary. For this case, we have that the adaptive combination aITKrM + aOMP achieved the best results with respect to all reported measures when compared to the non-adaptive combinations also for the best choices of S and K. Note that, pre-training the dictionaries was carried out on the kt-SENSE reconstructions obtained from $N_\theta = 3400$ radial spokes which do not contain severe artefacts. Further, for pre-training, we used a higher number of image patches and dictionary learning iterations than for learning the dictionary during the reconstruction. Therefore, by increasing the number of patches and iterations used to learn the dictionary online, one can most probably also expect an improvement of these results. In particular, because (as we will see in Subsection 5.4.3) aITKrM is approximately 10 times faster than K-SVD, allowing aITKrM to take the same amount of time as K-SVD, it is possible to surpass K-SVD + OMP also in terms of PSNR and to obtain the same NRMSE, for the case where the dictionary is learned during the reconstruction, see Table 5.3.

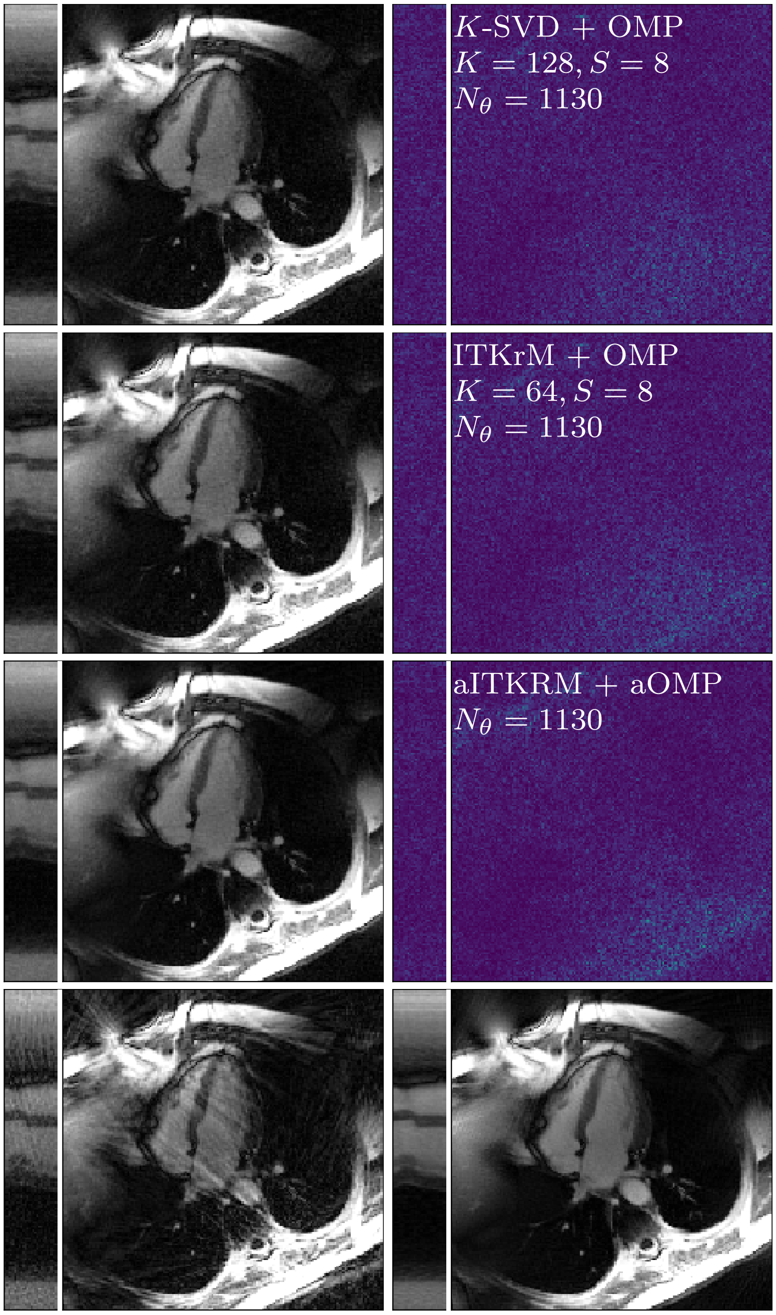


Figure 5.1: Results obtained by the best combinations of S and K for different dictionary learning and sparse approximation algorithms and their corresponding point-wise error-images. First row: K-SVD + OMP for $K = 128, S = 8$, second row: ITKrM + OMP for $K = 64, S = 8$, third row: aITKrM + aOMP, fourth row: initial NUFFT-reconstruction from $N_\theta = 1130$ radial spokes (left) and the kt-SENSE reconstruction using $N_\theta = 3400$ radial spokes which served as ground truth for the retrospective k-space data-generation (right). While the noise-level is similar for all combinations of dictionary learning and sparse approximation algorithms, using aITKrM + aOMP does not require the tuning of S and K.

Table 5.1: Comparison of the performance of different algorithms for dictionary learning and sparse coding used in the reconstruction. Using aITKrM + aOMP yields similar or better results compared to the ones obtained with the best combinations of S and K for the non-adaptive algorithms K-SVD + OMP and ITKrM + OMP.

	Ψ Learned during Reconstruction										
	Non-Adaptive										**Adaptive**
DL	**K-SVD**					**ITKrM**					**aITKrM**
SC	**OMP**					**OMP**					**aOMP**
K	64	128	128	128	256	64	128	128	128	256	ad.
S	8	16	8	4	8	8	16	8	4	8	ad.
PSNR	43.845	43.870	**44.538**	44.354	42.998	42.595	40.825	43.017	43.628	42.497	*44.491*
NRMSE	0.0681	0.068	**0.062**	0.064	0.075	0.079	0.096	0.074	0.069	0.079	*0.063*
SSIM	0.687	0.671	0.692	*0.710*	0.656	0.663	0.604	0.657	0.698	0.65	**0.734**
	Pre-Trained Ψ										
	Non-Adaptive										**Adaptive**
DL	**K-SVD**					**ITKrM**					**aITKrM**
SC	**OMP**					**OMP**					**aOMP**
K	64	128	128	128	256	64	128	128	128	256	ad.
S	8	16	8	4	8	8	16	8	4	8	ad.
PSNR	45.062	44.856	*45.205*	44.594	44.936	44.667	43.117	44.483	44.009	43.996	**45.314**
NRMSE	0.059	0.060	*0.058*	0.062	0.060	0.062	0.073	0.063	0.066	0.067	**0.057**
SSIM	0.703	0.684	0.699	*0.714*	0.691	0.696	0.645	0.687	0.709	0.675	**0.738**

5.4.2 Convergence behaviour

For assessing the convergence speed of the reconstruction algorithms, we tracked the different measures used for the evaluation of the performance of the reconstruction algorithms during the iterative reconstruction. Figure 5.2 shows the mean PSNR, NRMSE and SSIM averaged over the different images. Quite consistently, it can be observed that the reconstruction using the adaptive combinations aITKrM + aOMP surpassed the non-adaptive dictionary learning and sparse coding combinations at early iterates with respect to all measures and tended to let the curves flatten out earlier than the non-adaptive counterparts. This could be particularly well observed for the case of NRMSE and PSNR and held true for all scenarios with different S. ITKrM + OMP with $S = 16$ revealed a semi-convergent type of behaviour which can be attributed to the fact that $S = 16$ is too high for ITKrM in the presence of noise in k-space. This also shows that the choice of S can have a high impact on the reconstruction.

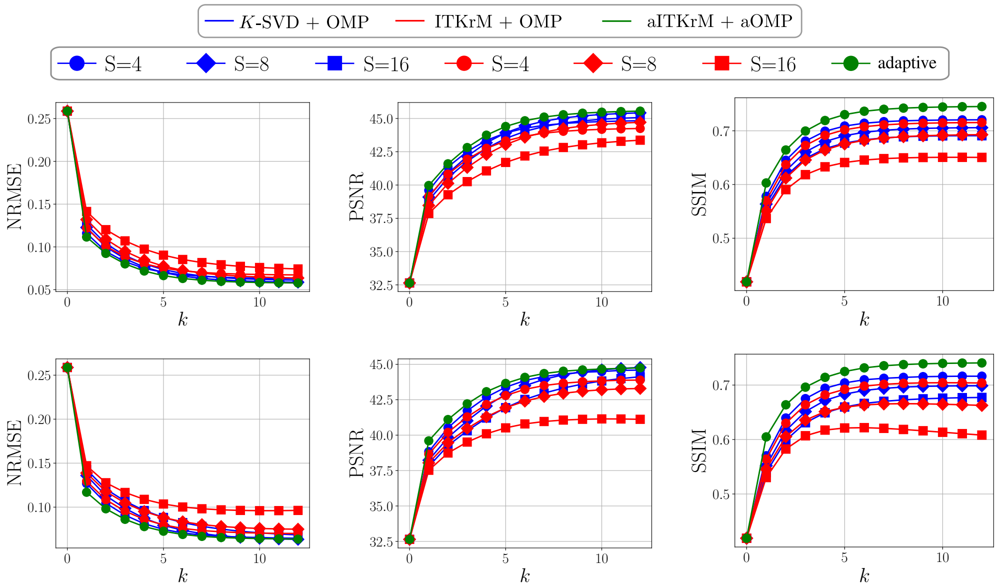


Figure 5.2: Convergence behaviour of the reconstruction scheme for solving (P1) (first row) and for solving (P2) (second row) using different combinations of dictionary learning and sparse approximation algorithms. The combination of aITKrM + aOMP yields better or equally good results compared to the non-adaptive combinations with respect to all measures, for solving (P1) and (P2). The images show the respective average measure over the iterations.

5.4.3 Reconstruction times

Here, we report the times for the different components of the dictionary learning-based reconstruction algorithms. The components which significantly contributed to the relatively high reconstruction times were the dictionary learning and sparse approximation algorithms and the PCG method which is needed to obtain an approximate solution of (5.4). Obviously, the latter was constant for the three different combinations of dictionary learning and sparse coding. Table 5.2 lists the average time needed for dictionary learning and sparse coding for each ADMM iteration. Therefore, the overall time needed for a specific component can be obtained by multiplying the respective time by the number of ADMM iterations T.

Table 5.2: Comparison of dictionary learning (DL) and sparse coding (SC) in terms of computational times in seconds for one ADMM iteration for solving problem (P2). We see that for the combination aITKrM + aOMP, the required computational time is the lowest for dictionary learning as well as for sparse coding. In each iteration the dictionary was learned on $N = 20\,000$ patches, for $n_{\mathrm{DL}} = 10$ iterations of the respective dictionary learning algorithm.

DL and SC	**Sparsity Level**	**DL / SC Time**
K-SVD + OMP	$S = 16$	71 / 849
	$S = 8$	69 / 415
	$S = 4$	69 / 206
ITKrM + OMP	$S = 16$	9 / 824
	$S = 8$	8 / 412
	$S = 4$	8 / 205
aITKrM + aOMP	adaptive	**7 / 149**

We see that K-SVD was the slowest dictionary learning algorithm and took approximately 69-71 seconds for one single ADMM iteration. ITKrM was considerably faster and took only between 8-9 seconds whereas its adaptive version was the fastest and took only around 7 seconds (including also the time which was needed to estimate the sparsity level and replacing coherent and unused atoms). For sparse coding, we see that for OMP the chosen sparsity level obviously had an impact on the required computational time and took 824-849 seconds for $S = 16$, 412-415 seconds for $S = 8$ and 205-206 seconds for $S = 4$. Our adaptive version aOMP was even faster as OMP for the lowest choice of $S = 4$ and required about 149 seconds.

Figure 5.3 shows a diagram representing the overall time for the respective component of the reconstruction algorithm. From the bars we can see the time which each component took relative to the total reconstruction time. First, we see that for the non-adaptive experiments, the time needed for the sparse approximation of all patches constitutes the computational bottleneck of the method when S is chosen too high, i.e. $S = 16$. Second, we see that, as expected, ITKrM was able to substantially reduce the computational time compared to K-SVD. However, the gain in terms of acceleration was negligible when putting it in relation to the overall time because OMP still remains the computational overhead for $S = 16$. The last bar of the graph shows that first, by employing aITKrM, the time needed to learn the dictionary still amounted to approximately the same as for ITKrM, and second, in this configuration, the time needed for sparse coding was clearly reduced and approximately corresponds to the one for OMP with $S = 4$.

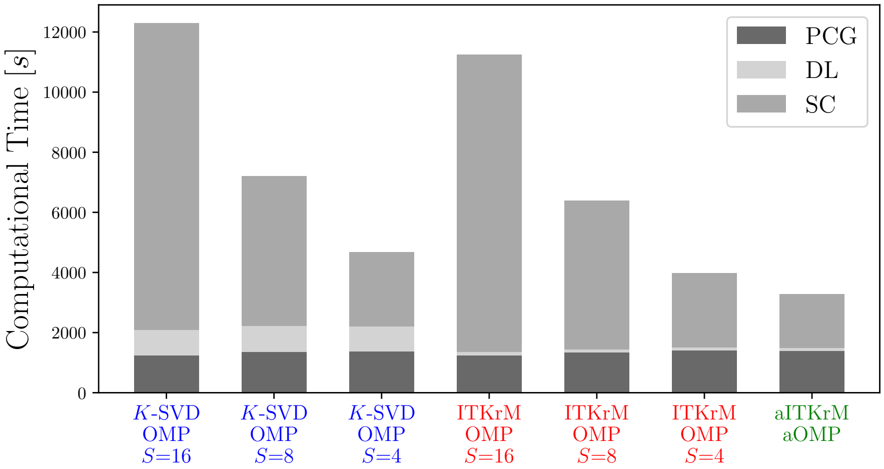


Figure 5.3: Reconstruction times grouped by components for different combinations of dictionary learning and sparse approximation algorithms for solving problem (P2) with $K = 128$. When solving (P1), the times needed for PCG and sparse coding remain similar, while the time for learning the dictionary Ψ can be neglected since it is assumed to be given a-priori.

5.4.4 Sensitivity with respect to $\mu_{\max}$ and M

For aITKrM we have to choose two input parameters: the maximal allowed coherence of the dictionary $\mu_{\max}$ and the minimal number of observations M. Here, we tested the stability of the reconstruction with respect to a variation of $\mu_{\max}$ and M. For this, we reconstructed all $n = 15$ cine MR image sequences with different combinations of them, namely for $\mu_{\max} = 0.5, 0.7$ and 0.9 and for $M = d, 2d$ and $2d \log d$. These experiments were carried out when solving problem (P2), i.e. where the dictionary Ψ is learned during the reconstruction. In order to ensure a better convergence behaviour of aITKrM during each learning stage, during the reconstruction, the number of iterations was set to $n_{\text{DL}} = 100$ and the number of patches used for learning Ψ was set to $N = 40\,000$. Note that by doing so, the total computational time required by aITKrM is less than for K-SVD, amounting to approximately 200 seconds for each learning stage. The obtained results can be found in Table 5.3. From the results we see that they are relatively stable with respect to the different choices of $\mu_{\max}$ and M.

5.4.5 Experiments using real k-space data

In the following, we tested the reconstruction algorithm with the different combinations of dictionary learning and sparse approximation algorithms by using the real k-space

Table 5.3: Variation of the pre-defined maximal allowed coherence $\mu_{\max}$ of the dictionary and the minimal number of observations M. For the experiments, the dictionary Ψ was learned during the reconstruction. For each iteration the dictionary was learned on $N = 40\,000$ patches using aITKrM with $n_{\mathrm{DL}} = 100$. We see that the results are relatively stable with respect to different choices of $\mu_{\max}$ and M.

$\mu_{\max}$	0.5	0.5	0.5	0.7	0.7	0.7	0.9	0.9	0.9
M	d	$d \log d$	$2d \log d$	d	$d \log d$	$2d \log d$	d	$d \log d$	$2d \log d$
PSNR	44.030	44.026	44.288	44.682	44.117	43.010	44.669	44.664	44.459
NRMSE	0.067	0.067	0.065	0.062	0.066	0.076	0.062	0.062	0.063
SSIM	0.732	0.732	0.733	0.734	0.731	0.723	0.734	0.733	0.732

data acquired along $N_\theta = 1130$ radial trajectories obtained from the scanner and compared it to *kt*-SENSE using $N_\theta = 3400$ radial trajectories. Note that sampling k-space along $N_\theta = 3400$ spokes already corresponds to an undersampling factor of ~ 3 which is needed to perform the scan in a single breathhold. Further, the *kt*-SENSE reconstruction algorithm itself imposes prior information to regularise the inverse problem and therefore, the *kt*-SENSE reconstructions obtained from the $N_\theta = 3400$ radial spokes cannot be considered as ground truth images for this experiment. Therefore, we abstain from reporting quantitative measures as well as point-wise error images. A rigorous quality assessment would need to be performed with respect to predefined clinical features and a clinical application. However, since this is beyond the scope of this work, we only show an example of the reconstruction for the sake of completeness and to demonstrate the applicability of aITKrM and aOMP for real k-space data. Figure 5.4 shows an example of images reconstructed with the three different combinations of dictionary learning and sparse approximation algorithms. The first row in Figure 5.4 shows the results obtained with K-SVD + OMP and the second row with ITKrM + OMP for different sparsity levels S, respectively. In the third row, we have the initial NUFFT-reconstruction, the result obtained with aITKrM + aOMP as well as the *kt*-SENSE reconstruction using $N_\theta = 3400$ radial spokes. Visually, all methods performed similarly well, and K-SVD + OMP and ITKrM + OMP show a slightly higher noise level compared to aITKrM + aOMP, which is consistent with the results presented in Subsection 5.4.1. However, note again that the times needed to obtain the reconstructed images are substantially lower for aITKrM + aOMP and no a-priori choice of the hyper-parameters S and K was required.

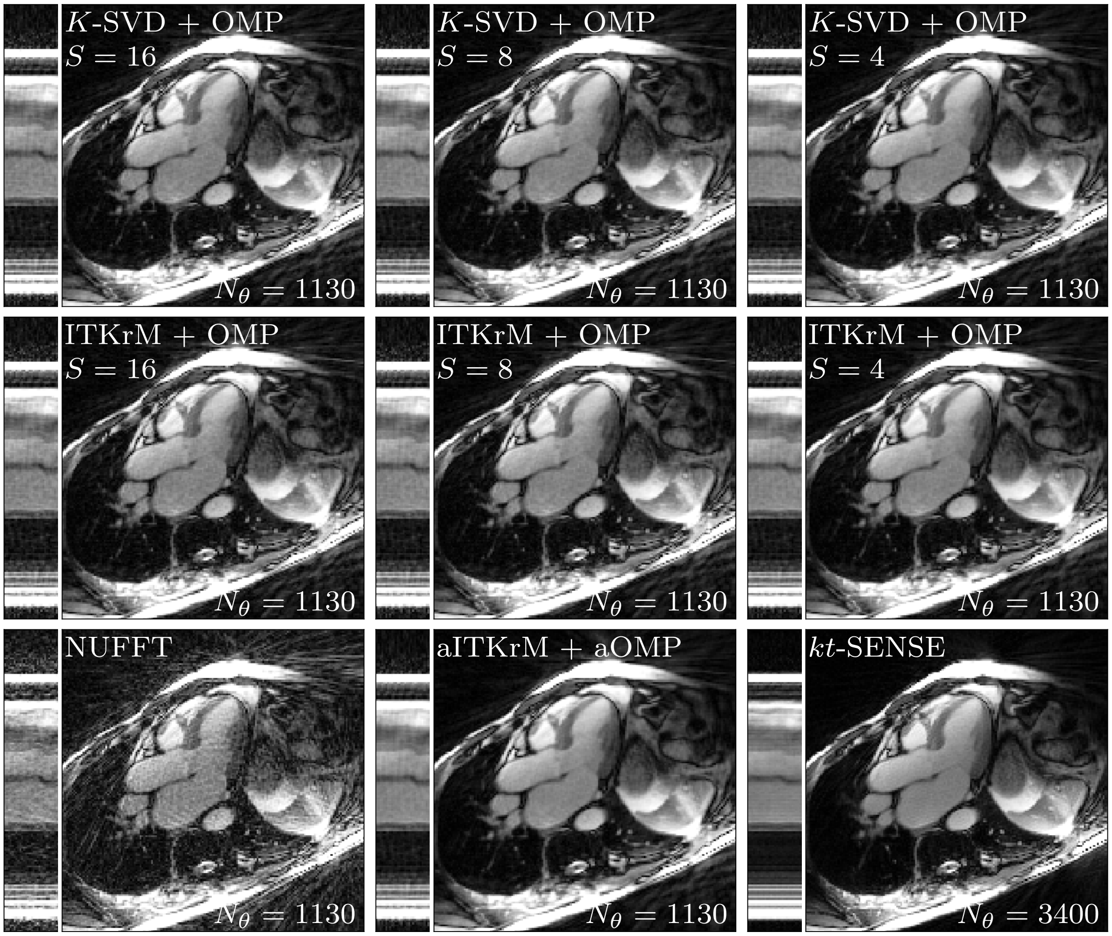


Figure 5.4: Results obtained from real k-space data obtained from the scanner measurements. Top row: K-SVD + OMP with $S = 16$, $S = 8$ and $S = 4$, mid row: ITKrM + OMP with $S = 16$, $S = 8$ and $S = 4$, third row: NUFFT-reconstruction using $N_\theta = 1130$ radial spokes, aITKrM + aOMP and kt-SENSE using $N_\theta = 3400$ radial spokes.

5.5 Discussion

We have seen that the adaptive versions of dictionary learning and sparse approximation algorithms given by aITKrM and aOMP provide valid alternatives to the well-established K-SVD algorithm and the non-adaptive sparse approximation algorithm OMP. In the following we discuss the advantages and limitations of the described algorithms in more detail.

5.5.1 Adaptive estimation of S and K

The major advantage of the combination aITKrM + aOMP is clearly that it is no longer necessary to choose the sparsity level S and the number of atoms K. This is important not only to make such algorithms more eligible for practical applications but also because a wrong choice of S and K can have a large impact on the computation time and the reconstruction quality. Intuitively speaking, within the sparse coding stage a too small choice of S leads to too smooth results with probably missing details while a too high choice of S results in a preservation of undersampling artefacts which we are trying to remove. Also, as the structure of an image varies from location to location, S should vary dependent on the considered image patch as well. Within the dictionary learning stage, in noisy situations a too high choice of S can cause the atoms to be disturbed by adding noise.

Moreover, the optimal number of atoms K is also data-dependent. In particular, for dictionaries learned on images containing more structure, a larger K is needed than for fairly smooth ones. Further, the optimal size of the dictionary was also shown to be dependent on the noise level of a corrupted image, i.e. the more noise, the smaller K should be chosen, [49]. These observations suggest that a global choice of S and K cannot be optimal, disregarding the fact that they are not known and can only be guessed. Using aITKrM and aOMP, S and K are adaptively chosen based on the texture of the current image estimate during the learning of the dictionary as well as during the sparse approximation step. Intuitively, at early iterations of the iterative reconstruction, a stronger regularisation of the image estimate is required in order to reduce the artefacts. At later iterations, where the current image estimate contains less noise and artefacts, a higher S and K are required to be able to represent fine anatomic details.

In order to illustrate how the dictionary size K as well as the sparsity level of an image varies during the iterative reconstruction, we conducted some additional experiments. For that, while solving problem (P2), meaning when the dictionary Ψ is learned during the reconstruction, we tracked the average dictionary size K estimated by aITKrM during the dictionary learning stage as well as the estimated sparsity level of each patch during the subsequent sparse coding stage using aOMP. As previously, the number of patches used to learn the dictionary was set to $N = 40\,000$ and the number of iterations was set to $n_{\mathrm{DL}} = 100$. The maximal allowed coherence of the dictionary and the minimal number observations were set to $\mu_{\max} = 0.7$ and $M = d \log d$, respectively. The results are shown in Figure 5.5 and Figure 5.6.

Figure 5.5 shows the average dictionary size K estimated by aITKrM during the reconstruction. We see that using aITKrM, the estimated number of atoms K needed to optimally represent the patches of the current image estimate tends to first decrease and then increase over the iterations.

In Figure 5.6 (a1) and (b1), the real and imaginary part of the NUFFT-reconstruction

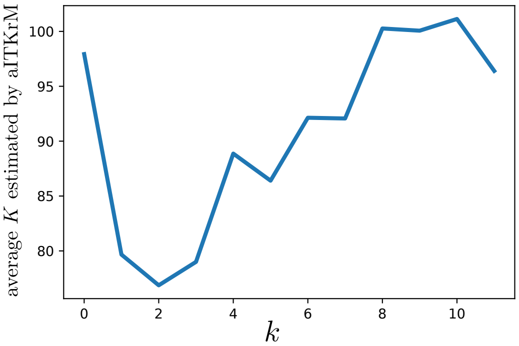


Figure 5.5: Average dictionary size K estimated by aITKrM during the iterative reconstruction.

$\mathbf{y}_U$ are displayed. In (c1) and (d1), we can see the corresponding patch-wise approximated images using aOMP and a dictionary learned by aITKrM. Figure 5.6 (e1) and (f1) show the estimated sparsity levels at various locations in the image. The second panel of Figure 5.6 shows the same images at the penultimate iteration $T = 11$ of the reconstruction. As we can see in (e2) and (f2), the average estimated sparsity level S is significantly higher than for the NUFFT-reconstruction, especially in the regions of the image which contain the patient's anatomy. In contrast, regions not containing the patient's anatomy but only background are sparsely approximated using a lower S.

This demonstrates that for the specific task of iterative image reconstruction, the optimal sparsity level S of a patch first of all depends on the needed complexity to represent relevant features and second, might change during the reconstruction. Further, in Subsection 5.4.3, we have observed that choosing S too high clearly has significant impact on the computational time and at the same time does not necessarily increase the reconstruction quality.

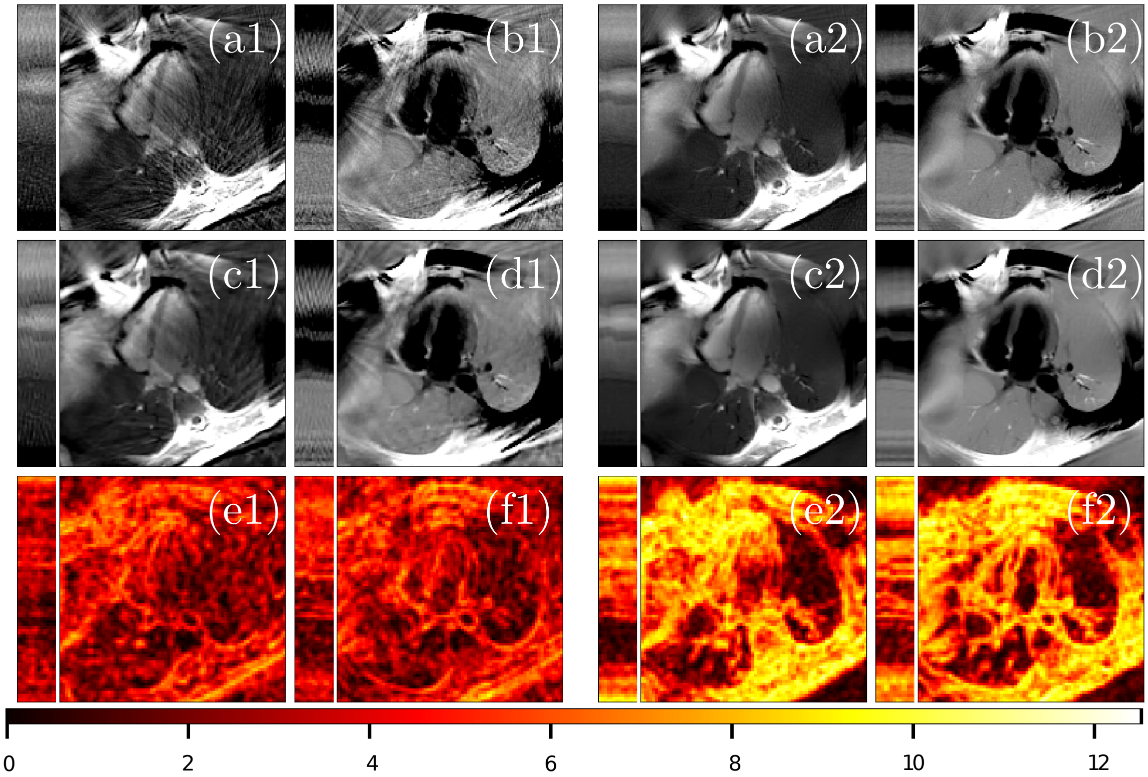


Figure 5.6: Estimated sparsity level at different stages during the iterative reconstruction for solving (P2), where Ψ is learned during the reconstruction. Left panel: (a1) and (b1) - real and imaginary part of the initial NUFFT-reconstruction $\mathbf{y}_U$, (c1) and (d1) - the correspondent patch-wise sparse approximations using aITKrM + aOMP, (e1) and (f1) - the estimated sparsity levels of the image-patches at various locations. Right panel: (a2) and (b2) - real and imaginary part of the twelfth iterate obtained by using aITKrM + aOMP, (c2) and (d2) - the correspondent patch-wise sparse approximations using aITKrM + aOMP, (e2) and (f2) - the estimated sparsity levels of the image-patches at various locations. The average sparsity level S is therefore lower at early iterates in the reconstruction and higher at later iterates.

In Figure 5.7 we see an example of eight atoms out of the dictionaries learned by the respective dictionary learning algorithms. The atoms of the dictionaries shown in the figure were learned on a set of patches extracted from the initial NUFFT-reconstruction $\mathbf{y}_U$ (first row) and from the penultimate image estimate of the reconstruction (second row). We can see that the dictionaries learned by the non-adaptive dictionary learning algorithms with $S = 16$ tend to inherently contain quite a large portion of noise in the atoms which, on the other hand, is almost not present in the atoms learned by aITKrM. This observation is consistent with the theory discussed in Subsection 4.3.1, for the case where the sparsity level S is overestimated and suggests that $S = 16$ is a far too high choice. The fact that S and K no longer need to be chosen a-priori could highly facilitate a possible application of the reconstruction algorithm in the clinical routine, where standardised acquisition and reconstruction protocols have to be used. Further, as we have seen in the examples shown in Subsection 5.3, the S- and K-adaptivity achieves competitive results compared to K-SVD + OMP and additionally reduces the required reconstruction times.

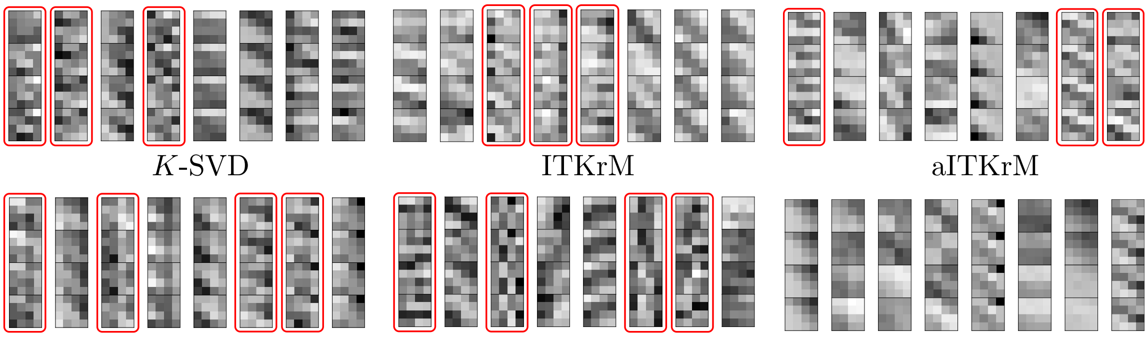


Figure 5.7: Examples of eight three-dimensional atoms (un-stacked along the time dimension) of the dictionaries learned by K-SVD (left), ITKrM (mid) and aITKrM (right). The dictionaries were learned on 3D patches extracted from the initial NUFFT-reconstruction $\mathbf{y}_U$ (first row) and the penultimate image estimate (second row). For K-SVD and ITKrM, the sparsity level was $S = 16$. Since $S = 16$ is relatively high, some of the atoms obtained by K-SVD and ITKrM contain a relatively large portion of noise (especially the ones marked in red). For aITKRM, the atoms seem to be considerably more stable, in particular at the penultimate iteration (second row). Note that the constant atom is not shown in the images.

5.5.2 Limitations

A possible limitation is that the thresholds chosen for the algorithms underlie the theoretical consideration of Gaussian and sub-Gaussian noise which might not be true in general. However, sampling along radial trajectories is known to represent an incoherent sampling pattern with noise-like properties and similar or even better results could be probably obtained by using Compressed-Sensing Cartesian schemes [40].
As all iterative reconstruction methods which employ a-priori knowledge expressed as a penalty term, the dictionary learning-based regularisation method requires to choose the regularisation parameter λ. However, quite some work has been dedicated on how to adaptively choose the parameter λ as well, see e.g. [12, 37], which might be incorporated in the reconstruction algorithm using aITKrM + aOMP.

5.5.3 Dictionary learning vs. deep learning

Especially with the emergence of deep learning-based methods for the regularisation of inverse problems in medical image reconstruction, see for example [56], [26], [69], [68], [27], [31], one can ask whether dictionary learning-based regularisation has nowadays become an obsolete regularisation method. Nevertheless, without a doubt, the black-box character of deep learning remains an open issue which still needs to be properly addressed, in particular when used for medical imaging applications [41]. In fact, deep learning-based methods for image reconstruction have recently been reported to be affected by instabilities [4].

In contrast, dictionary learning has a longer tradition and stands on a more solid mathematical foundation with well-understood theory. Therefore, opting for dictionary learning and sparse coding as regularisation methods offers the possibility to employ Machine Learning-based methods with a more profound theoretical understanding. Further, note that often a reason why deep learning-based methods are also favoured over other regularisation methods is their fast application using appropriate libraries and GPUs. However, for large-scale problems (as the one considered in this work), even if obtaining a regularised image with a neural network is fast, in order to increase data-consistency of the solution, an appropriate functional should be subsequently minimised, see e.g. [32]. Therefore, the overall required computational time for this type of approaches is mainly dependent on the implementation of the iterative solver, i.e. on the implementation of the forward/adjoint models. Note that in our proposed method, the time needed for dictionary learning and sparse coding amounts to approximately the same as the time needed for PCG. Thus, the longer computational time compared to deep learning-based methods seems to be an acceptable price to pay for being able to use a theoretically well-founded regularisation method which also does not require the a-priori choice of the hyper-parameters S and K.

5.5.4 Reconstruction quality

The achieved image quality using aITKrM + aOMP is comparable with the one achieved using the standard combination K-SVD + OMP with the best reported choices of the sparsity level as can be seen in Figure 5.1 and Table 5.1. The performed experiments reveal that for K-SVD, choosing S too high impairs image quality compared to a lower choice of S. This effect is even clearer for ITKrM, where a too high S is known to disturb atoms in the dictionary, especially in the presence of noise. Moreover, in Figure 5.3 we have seen that overestimating S leads to a substantial increase of computational time. From these experiments we can further conclude that the choice of S is non-trivial. Also, relying on the choice of hyper-parameters suggested in the literature might not be optimal, as the reported parameters are always data- and problem dependent and usually adapted to a specific task. This observation makes the S- and K-adaptivity a particularly interesting feature of the combination aITKrM + aOMP from a practical point of view. First of all, it is possible to reduce the computation time and second, it is possible further improve the reconstruction quality. Although aITKrM requires the choice of the maximal allowed coherence $\mu_{\max}$ of the dictionary and the minimal number of observations M, we have seen that the combination of aITKrM and aOMP was relatively robust with respect to the latter. Note that, while S and K depend on the type of signals, M only depends on the number of available training signals per iteration.

5.5.5 Reconstruction times

Learning a dictionary with aITKrM instead of K-SVD leads to an acceleration factor of approximately 10 which is useful when the dictionary is learned during the reconstruction. The reason is that the computationally most expensive component of K-SVD is OMP, where aITKrM in contrast only requires the faster thresholding. More importantly, using aOMP has the potential of highly reducing the time needed for the sparse approximation of all patches since, instead of using a (as we have seen, potentially too high) global sparsity level S, it is adaptively chosen according to the considered patch-example.

Summarising this chapter, we have investigated the application of aITKrM and aOMP for the task of cine MR image reconstruction. We have shown their competitiveness and advantages compared to the well-established K-SVD and OMP algorithms. While most methods employing dictionary learning and sparse coding for the regularisation of image reconstruction in MR use a global sparsity level S for learning the dictionary as well as for sparsely approximating the image patches, we have seen that this can never be optimal. Further, S and the number of atoms K to be used are usually determined by computationally expensive hyper-parameter searches. Using aITKrM and aOMP, S and K are adaptively chosen dependent on the texture of the currently considered image estimates. As we have seen, aOMP provides appropriate estimates of S for the sparse approximation of the patches and by this, a more efficient regularisation is achieved. This also results in a significant acceleration of the regularisation step, especially when compared to the case for standard a-priori choices of S and K.

From the comparison of the reconstruction times we have seen that aITKrM and ITKrM were significantly faster than K-SVD. In particular, as thresholding has a much lower computational complexity than OMP, we were able to substantially accelerate the dictionary learning stage of the iterative reconstruction. The computational lightness of thresholding however comes at the price of a much weaker performance compared to OMP. Interestingly, although K-SVD uses the much better sparse approximation algorithm OMP, all algorithms yielded similar results. Further, also in [61] it was shown that K-SVD and ITKrM perform equally well. To see if we can get some theoretical insights which shed light on the question whether in dictionary learning the performance of OMP is worth its cost, in the next chapter we will analyse the performance of OMP for input dictionaries that are affected by some perturbation and hence do not coincide with the signal generating dictionary.

Chapter 6

Orthogonal Matching Pursuit (OMP) with Perturbations

In this chapter we have a closer look at sparse approximation algorithms. In particular, we present average case results for OMP in case where we do not have the signal generating dictionary but only a perturbed version of it. We provide recovery conditions for noiseless as well as noisy signals and further compare them with conditions obtained for thresholding. Finally we conduct various numerical experiments in order to illustrate our theoretical findings.

6.1 Limitations of Existing Results

In the previous chapter we have seen how well sparse approximation algorithms are suited for real-world applications. Besides their good practical performance, there exists also detailed theory that analyses their worst case or average case performance, see e.g. [62, 65, 60, 24, 66]. However, these results are usually based on the assumption that the signal generating dictionary is given. This means, the dictionary used in the signal model is also assumed to be given as input parameter for the corresponding sparse approximation algorithm. Indeed, such assumptions may not hold true in all situations. In practical applications we in general have at best a good approximation of this dictionary. Moreover, sparse approximation algorithms are also used within dictionary learning algorithms where the learned dictionary can be completely different from the generating dictionary, especially in the first iterations. In this chapter we want to bridge the gap between theory and practice and provide recovery conditions for OMP for the case where we do not have the signal generating dictionary. In particular, we provide average case results for OMP where the given input dictionary is only a perturbed version of the generating dictionary, for noiseless as well as noisy signals. Comparing the theoretical guarantees for OMP and thresholding, we will see that in presence of perturbations both conditions contain a term that limits the range of parameters for which they perform well. In the perturbation-free case, on the other hand, this term only occurs in the recovery conditions of thresholding.

This chapter is organised as follows. After introducing the signal model and the dictionaries on which our results are based, in Section 6.2 we derive recovery conditions for OMP in case of noiseless signals. The results concerning the noisy case are presented in Section 6.3. In Section 6.4 we provide recovery conditions for thresholding and compare them with the ones of OMP. In order to illustrate the obtained results we conduct various numerical experiments in Section 6.5 and conclude the chapter in Section 6.6.

6.1.1 Signal model and dictionaries

Before we start, we have to introduce the signal model on which our results are based and explain how we generate a perturbation on our dictionary. Note that, this is essentially the same as in Subsection 3.1.1 and Section 1.2 but to simplify the proofs we use a slightly different notation.

Signal model

Given a $d \times K$ dictionary Φ, we assume that our signals are generated as

$$y = \Phi_{p(I)} x_I := \sum_{i \in I} \phi_{p(i)} x_i \qquad \text{with} \qquad x_i = \sigma_i c_i, \tag{6.1}$$

where $I = \{1, \dots S\}$, $(\sigma_i)_i$ forms a Rademacher sequence, the coefficients c_i are non-increasing, meaning, we have that $c_1 \geq c_2 \geq \dots \geq c_S \geq 0$ and p is some permutation of $\{1, \dots K\}$ chosen uniformly at random among all permutations. For convenience we write for any index set J, $\bar{J} = p(J)$ and hence, $\Phi_{p(I)} x_I = \Phi_{\bar{I}} x_I$.

In order to prove success of OMP we use similar ideas as already used in [60]. In particular, in order to get useful average case results, we need coefficients exhibiting some decay. In this case, we get a natural order in which atoms are more likely to be picked, atoms corresponding to larger coefficients, before other atoms corresponding to smaller coefficients.
For our analysis we group the indices of the S non-zero coefficients into s slots depending on the coefficient size. We take $\beta \in (0,1)$ and define the map

$$b : \{1, \dots, S\} \to \{1, \dots, s\} \tag{6.2}$$

$$b(i) = j \Leftrightarrow c_i \in (c_1 \beta^j, c_1 \beta^{j-1}] \tag{6.3}$$

and the slots U_j via

$$U_j = b^{-1}(\{j\}). \tag{6.4}$$

To assign every non-zero coefficient index to a slot U_j the number of slots needs to satisfy $c_1 \beta^s < c_S \leq c_1 \beta^{s-1}$, meaning $s = \lceil \frac{\log(c_S/c_1)}{\log \beta} \rceil + 1$. Note that at most S of these slots are non-empty.

Let i be the smallest still missing index, meaning the index of the largest still missing coefficient. Based on i we define the following disjoint sets,

$$\begin{aligned} A_i &:= \bigcup_{\ell=1}^{b(i)-1} U_\ell \quad \text{with} \quad A_1 = \emptyset, \\ M_i &:= U_{b(i)}, \\ N_i &:= U_{b(i)+1}, \\ R_i &:= \mathbb{S}^c \cup \bigcup_{\ell=b(i)+1}^{s} U_\ell. \end{aligned}$$

This means, A_i contains all indices corresponding to coefficients which are larger than the largest still missing coefficient c_i. The set M_i contains i and all indices of coefficients which are in the same slot as c_i, meaning, which are approximately of the size of c_i. The set N_i comprises all indices of coefficients in the subsequent slot, which are large enough to be picked but certainly smaller than the largest missing one. R_i contains the indices of all remaining coefficients, which are small and hence, very unlikely to be picked before the ones in M_i.

Dictionaries

Given the generating dictionary $\Phi = (\phi_1, \dots, \phi_K)$, we use the same decomposition as in Section 1.2 to model a perturbation of it, $\Psi = (\psi_1, \dots, \psi_K)$, but with some additional notation. Remember that the distance between Φ and Ψ is defined as $d(\Phi, \Psi) = \max_k \|\phi_k - \psi_k\|_2 = \max_k \varepsilon_k = \varepsilon$. So we can find unit vectors z_k with $\langle z_k, \phi_k \rangle = 0$ such that

$$\psi_k = \alpha_k \phi_k - \omega_k z_k \quad \text{for} \quad \alpha_k := 1 - \frac{\varepsilon_k^2}{2} \quad \text{and} \quad \omega_k := \left(\varepsilon_k^2 - \frac{\varepsilon_k^4}{4} \right)^{\frac{1}{2}}. \tag{6.5}$$

Conversely we can also decompose the generating atoms as

$$\phi_k = \gamma_k \psi_k + \lambda_k z_k \quad \text{for} \quad \gamma_k := \alpha_k^{-1} = \frac{2}{2 - \varepsilon_k^2} \quad \text{and} \quad \lambda_k := \frac{\omega_k}{\alpha_k} = \gamma_k \omega_k. \tag{6.6}$$

Let $Z = (z_1, \dots, z_K)$ denote the perturbation dictionary, collecting the perturbation vectors z_k as its columns, and define the diagonal matrices $A = \mathrm{diag}((\alpha_k)_k)$, $W = \mathrm{diag}((\omega_k)_k)$, $\Gamma = \mathrm{diag}((\gamma_k)_k)$ and $\Lambda = \mathrm{diag}((\lambda_k)_k)$, with the corresponding constants on their diagonal. Hence, the perturbed dictionary Ψ is given by

$$\Psi = \Phi A - ZW, \tag{6.7}$$

and the generating dictionary by

$$\Phi = \Psi A^{-1} + ZWA^{-1} =: \Psi \Gamma + Z \Lambda. \tag{6.8}$$

We also define $\gamma = \max_k \gamma_k$, $\gamma_{\min} = \min_k \gamma_k$, $\varepsilon_{\min} = \min_k \|\phi_k - \psi_k\|_2$ and $\nu_Z = \max_{i,j} |\langle \psi_i, z_j \rangle|$.

Using these definitions, in the following sections we provide recovery conditions for OMP in the case where we do not have the generating dictionary Φ but only a perturbed version of it, Ψ, for noiseless perfectly S-sparse signals as well as signals which are contaminated with noise.

6.2 Recovery Conditions for Noiseless Signals

We start with deriving recovery conditions for OMP for the simple case of noiseless signals. In particular, assuming that the signals follow the model in (6.1), we provide conditions ensuring that with high probability, OMP recovers the generating support.

Theorem 6.1. *Assume that the signals follow the model in (6.1) with coefficients grouped by their magnitude into s slots U_ℓ as defined in (6.4). Let t denote the maximal number of elements within a slot and Ψ some perturbation of the generating dictionary Φ as defined in (6.7) with $\varepsilon \leq 1$. Further, assume that*

$$\mu(\Psi) \leq \frac{1}{4m \log K} \quad \text{and} \quad S \leq \frac{K}{16me^2\|\Psi\|_{2,2}^2 \log K}.$$

Then OMP will recover the full support, except with probability $K(2sK^{-2n} + 3K^{-n} + 216K^{-m})$, as long as

$$1 - \frac{\gamma}{\gamma_{min}}\beta \geq 4\,\varepsilon \cdot \frac{\gamma}{\gamma_{min}}\sqrt{n \log K} \cdot \max\left\{\frac{2\nu_Z}{\beta^s}\sqrt{n \log K}, \frac{\|c_I\|_2}{c_1\beta^s}\sqrt{\frac{\|Z\|_{2,2}^2}{K-S}}\right\} \cdot \xi$$
$$+ 4\mu(\Psi) \cdot \frac{\gamma}{\gamma_{min}}\left(t + \sqrt{nt \log K} \cdot \left(\frac{\beta^2}{1-\beta^2}\right)^{\frac{1}{2}}\right) \cdot \xi, \tag{6.9}$$

(a) with $\xi = 1 + 4\mu(\Psi)\sqrt{2Sn \log K}$ for $S\|\Psi\|_{2,2}^2/K \leq 4\mu(\Psi)^2 n \log K$,

(b) and $\xi = 1 + 2S \cdot \sqrt{\frac{2\|\Psi\|_{2,2}^2}{K}}$ for $S\|\Psi\|_{2,2}^2/K \geq 4\mu(\Psi)^2 n \log K$,

where $\varepsilon = \max_k \|\phi_k - \psi_k\|_2$, $\gamma = \frac{2}{2-\varepsilon^2}$, $\gamma_{min} = \frac{2}{2-\varepsilon_{min}^2}$ with $\varepsilon_{min} = \min_k \|\phi_k - \psi_k\|_2$ and $\nu_Z = \max_{i,j} |\langle \psi_i, z_j\rangle|$.

Before we prove the theorem we would like to say a few words about the result. The theorem consists of two parts, however, as in general $4\mu(\Psi)^2 n \log K \leq S\|\Psi\|_{2,2}^2/K$, we restrict our discussion to this case. The other case follows directly by inserting the corresponding bounds.

The condition on the coherence of the perturbed dictionary $\mu(\Psi)$ and the sparsity level S ensures that for a randomly chosen subset $\bar{I}$ (permutation p) we have $\delta_{\bar{I}}(\Psi) \leq \frac{1}{2}$ except with probability $216K^{1-m}$. For $\varepsilon \leq 1$, the constant $\frac{\gamma}{\gamma_{\min}} = \frac{2-\varepsilon_{\min}}{2-\varepsilon}$ is bounded by $1 \leq \frac{\gamma}{\gamma_{\min}} \leq 2$ and hence, at worst we have an additional factor 2 on the right hand side. In order to get also a feeling for the remaining constants, assume that the generating dictionary Φ is well-behaved with coherence $\mu(\Phi) \approx \frac{1}{\sqrt{d}}$ and operator norm $\|\Phi\|_{2,2}^2 \approx \frac{K}{d}$. For perturbed atoms ψ_k following the definition in (6.5) with perturbation vectors z_k drawn uniformly at random from the unit sphere, with high probability, for all $j \neq k$, we have

$$|\langle \phi_k, z_j \rangle| \lesssim \sqrt{\log K/d} \quad \text{and} \quad |\langle z_k, z_j \rangle| \lesssim \sqrt{\log K/d},$$

and therefore, $\mu(\Psi) \lesssim \sqrt{4\log K/d}$. For the operator norm of the perturbation dictionary Z we have that with high probability $\|Z\|_{2,2} \lesssim \sqrt{\log K}$, [64], and hence, $\|\Psi\|_{2,2} \lesssim \sqrt{K/d} + \sqrt{\log K}$. For $\nu_Z = \max_{i,j} |\langle \psi_i, z_j \rangle|$, we have $\nu_Z \lesssim \sqrt{\log K/d}$ for $i \neq j$ and $\nu_Z \lesssim \varepsilon$ for $i = j$. Considering the special case where all atoms are equally perturbed, meaning $\varepsilon_{\min} \approx \varepsilon$, the condition ensuring support recovery, except with probability $K(2sK^{-2n} + 3K^{-n} + 216K^{-m})$, is given by

$$\begin{aligned}\frac{1-\beta}{4} \gtrsim \varepsilon \frac{n\log K}{\beta^s} \cdot \max\left\{2\varepsilon, \sqrt{\frac{4\log K}{d}}, \frac{\|c_I\|_2}{c_1\sqrt{K}}\right\} \left(1 + \frac{4S}{\sqrt{d}}\right) \\ + \sqrt{\frac{4\log K}{d}} \left(t + \sqrt{nt\log K} \cdot \left(\frac{\beta^2}{1-\beta^2}\right)^{\frac{1}{2}}\right) \left(1 + \frac{4S}{\sqrt{d}}\right). \qquad (6.10)\end{aligned}$$

Considering the condition in (6.10), we immediately observe that while a strong decay of the signal coefficients (and hence β small) is beneficial to have the left hand side large as well as the second term on the right hand side small, the opposite holds true for the first term on the right hand side. In particular, for β small, $1/\beta^s$ grows very fast. Especially for larger distances between the generating and the perturbed dictionary, meaning larger ε, this term is becoming increasingly dominant, thus decreasing the success rate of OMP. To confirm these observations, we conduct several numerical experiments with signals where the signal coefficients form a geometric sequence in Section 6.5.

Going back to theory, in case where we have the generating dictionary Φ, meaning $\varepsilon = 0$, Theorem 6.1 (b) states that OMP recovers the full support, except with probability $K(2sK^{-2n} + 2K^{-n} + 216K^{-m})$, as long as

$$\left(t + \sqrt{nt\log K} \cdot \left(\frac{\beta^2}{1-\beta^2}\right)^{\frac{1}{2}}\right) \left(\mu(\Phi) + 2S\mu(\Phi) \cdot \sqrt{\frac{2\|\Phi\|_{2,2}^2}{K}}\right) \leq \frac{1-\beta}{4}. \qquad (6.11)$$

In order to get a feeling for this result, let us compare it with some existing results and note that in general we have $\|\Phi\|_{2,2}/\sqrt{K} < \mu(\Phi)$. From the worst-case analysis in [65] we have that OMP succeeds as long as $2S\mu(\Phi) \leq 1$. In this case, the condition above holds trivially true. If conversely $2S\mu(\Phi) \geq 1$, we have that $\mu(\Phi) + 2S\mu(\Phi)^2 \leq 4S\mu(\Phi)^2$, and the condition in (6.11) says that OMP will recover the full support with high probability, as long as $St\mu(\Phi)^2 \lesssim (1-\beta)$ and $\sqrt{nt\log K} \cdot S\mu(\Phi)^2 \lesssim \frac{1-\beta}{\beta}\sqrt{1-\beta^2}$.

Considering the special case of coefficients forming a geometric sequence with decay factor $\alpha < 1$, and choosing $\beta = \alpha^k$ for $k \geq 1$, we have $t = k$. In this case, we need to have $\sqrt{n \log K} \cdot S\mu(\Phi)^2 \lesssim \frac{1-\alpha^k}{\alpha^k}\sqrt{(1-\alpha^{2k})/k}$ and $S\mu(\Phi)^2 \lesssim (1-\alpha^k)/k$. Choosing $t = 1$, we have $\beta = \alpha$ and the condition ensuring support recovery essentially says that we need to have $\sqrt{n \log K} \cdot S\mu(\Phi)^2 \lesssim \frac{1-\alpha}{\alpha}\sqrt{1-\alpha^2}$ and $S\mu(\Phi)^2 \lesssim 1-\alpha$, which is consistent with the result in [60].

In order to prove Theorem 6.1 we use a similar strategy to the one already used in [60]. In particular, in order to show that OMP adds only correct atoms within each iteration, for all possible sub-supports $\bar{J} \subseteq \bar{I} = p(I)$ and $r_{\bar{J}} = Q(\Psi_{\bar{J}})y$, we need to have

$$\max_{i \in I} \left|\langle \psi_{p(i)}, r_{\bar{J}} \rangle\right| \overset{!}{>} \max_{k \notin I} \left|\langle \psi_{p(k)}, r_{\bar{J}} \rangle\right|. \tag{6.12}$$

Since there are 2^S possible sub-supports one has to control, the idea in [60] to get useful average case results was to reduce this number of sub-supports by utilising the decay of the coefficients. This means, if we have decaying coefficients it is more likely that OMP picks atoms corresponding to larger coefficients before the ones corresponding to smaller coefficients, hence reducing the number of possible sub-supports. In particular, using the definition of the slots introduced in Section 6.1.1, for i the index of the largest still missing coefficient and for an appropriate choice of β, we always have $p(A_i) \subseteq \bar{J} \subseteq p(A_i \cup M_i \cup N_i)$.
In the following we show that OMP picks an index $p(j)$ with $j \in M_i \cup N_i$ rather than $p(k)$ with $k \in R_i$. This means, another correct atom is added and our support is again of the form $\bar{J} = p(J)$ with $J = A_{i_{\text{new}}} \cup G_{\text{new}} \subseteq A_{i_{\text{new}}} \cup M_{i_{\text{new}}} \cup N_{i_{\text{new}}} \subseteq I$, for i_{new} the index of the new largest still missing coefficient. Iterating this process, a sufficient condition for OMP to fully recover the support $\bar{I}$ is that for all sub-supports $\bar{J} = p(A_i \cup G) \subseteq \bar{I}$ with $G \subseteq M_i \cup N_i$, we have

$$\left|\langle \psi_{p(i)}, r_{\bar{J}} \rangle\right| > \max_{k \in R_i} \left|\langle \psi_{p(k)}, r_{\bar{J}} \rangle\right|, \tag{6.13}$$

where i denotes the index of the largest still missing coefficient.

The following proposition concerns the norm of a random subvector and is used to prove Theorem 6.1. Its proof can be found in Appendix A.2. We will also use Proposition 3.4 from Section 3.2 which we repeat for convenience.

Proposition 6.2. *Let $v \in \mathbb{R}^K$ be a vector, I a subset chosen uniformly at random among all subsets of size S, and v_I the restriction of v to the subset I, then for any $t \geq 0$,*

$$\mathbb{P}\left(\|v_I\|_2^2 \geq \tfrac{S}{K}\|v\|_2^2 + t\right) \leq \exp\left(-\frac{t^2}{2\left(\|v\|_\infty^2 t + \frac{S}{K}\|v\|_4^4\right)}\right). \tag{6.14}$$

Proposition 3.4. *Let $v \in \mathbb{R}^K$ be a vector, $I = (i_1, \dots, i_S)$ be a sequence of length S obtained by sampling from $\mathbb{K} = \{1, \dots, K\}$ without replacement, ε with values in $\{-1, 1\}^S$ a Rademacher vector independent from I and $c \in \mathbb{R}^S$ a scaling vector. Then for any $t \geq 0$,*

$$\mathbb{P}\left(\big| \sum_{k=1}^{S} c_k \varepsilon_k v_{i_k} \big| \geq t \right) \leq 2 \exp \left(- \frac{t^2}{2\big(\|c\|_\infty \|v\|_\infty t + \|c\|_2^2 \|v\|_2^2 / (K - S) \big)} \right).$$

Proof of Theorem 6.1. Throughout the proof of Theorem 6.1 we use the abbreviations $\bar{B} := \|\Psi\|_{2,2}^2$ and $\bar{\mu} := \max_{i \neq j} |\langle \psi_i, \psi_j \rangle|$. We also use the short hands already used in previous chapters, $Q(\Psi_J) = \mathbb{I}_d - P(\Psi_J)$ and $r_J = Q(\Psi_J) y$ for the residual based on some index set J.

Using the definition of the slots introduced in Section 6.1.1 with i the index of the largest still missing coefficient, for our analysis we assume that our current support is of the form $\bar{J} = p(J)$ with $J = A_i \cup G$ and $G = C \cup D$ with $C \subseteq M_i$ and $D \subseteq N_i$. Further, we define $C^c = M_i \setminus C$, $D^c = N_i \setminus D$ and $G^c = C^c \cup D^c$. For convenience, we again write $\bar{C}^c = p(C^c)$, $\bar{D}^c = p(D^c)$ and $\bar{G}^c = \bar{C}^c \cup \bar{D}^c$. To keep the indices under control we write R instead of R_i and hence, $\bar{R} := p(R) := p(R_i)$. Note that, for i the index corresponding to the largest still missing coefficient, we have $i \in C^c \subseteq G^c$.

Assuming our current support is of the form $\bar{J} = p(A_i \cup G) \subseteq \bar{I} = p(I)$, a sufficient condition ensuring that OMP picks another $j \in G^c$ and therefore, another correct (sub-) support, is given by

$$\big| \langle \psi_{p(i)}, r_{\bar{J}} \rangle \big| > \max_{k \in R} \big| \langle \psi_{p(k)}, r_{\bar{J}} \rangle \big| . \tag{6.15}$$

In order to show that the residual inner product with the atom indexed by $p(i)$ is larger than the one with atoms indexed by $p(k) \in \bar{R}$, we use the following decomposition of our signals, where $\bar{J}^c = \bar{G}^c \cup \bar{R}$,

$$y = \Phi_{\bar{I}} x_I = (\Psi\Gamma)_{\bar{I}} x_I + (\Phi - \Psi\Gamma)_{\bar{I}} x_I = (\Psi\Gamma)_{\bar{J}} x_J + (\Psi\Gamma)_{\bar{J}^c} x_{J^c} + (Z\Lambda)_{\bar{I}} x_I. \tag{6.16}$$

Hence, the residual is of the form

$$r_{\bar{J}} = Q(\Psi_{\bar{J}}) y = Q(\Psi_{\bar{J}})((\Psi\Gamma)_{\bar{J}^c} x_{J^c} + (Z\Lambda)_{\bar{I}} x_I), \tag{6.17}$$

and for the residual inner product with $\psi_{p(i)}$, we get

$$\begin{aligned} \big| \langle \psi_{p(i)}, r_{\bar{J}} \rangle \big| &= |\langle \psi_{p(i)}, Q(\Psi_{\bar{J}})(\Psi\Gamma)_{\bar{J}^c} x_{J^c} \rangle + \langle \psi_{p(i)}, Q(\Psi_{\bar{J}})(Z\Lambda)_{\bar{I}} x_I \rangle| \\ &\geq |\langle \psi_{p(i)}, Q(\Psi_{\bar{J}})(\Psi\Gamma)_{\bar{J}^c} x_{J^c} \rangle| - |\langle \psi_{p(i)}, Q(\Psi_{\bar{J}})(Z\Lambda)_{\bar{I}} x_I \rangle|, \end{aligned} \tag{6.18}$$

and for any atom $\psi_{p(k)}$ with $k \in R$,

$$\begin{aligned}\big|\langle \psi_{p(k)}, r_{\bar J}\rangle\big| &= |\langle \psi_{p(k)}, Q(\Psi_{\bar J})(\Psi\Gamma)_{\bar J^c} x_{J^c}\rangle + \langle \psi_{p(k)}, Q(\Psi_{\bar J})(Z\Lambda)_{\bar I} x_I\rangle \\ &\leq |\langle \psi_{p(k)}, Q(\Psi_{\bar J})(\Psi\Gamma)_{\bar J^c} x_{J^c}\rangle| + |\langle \psi_{p(k)}, Q(\Psi_{\bar J})(Z\Lambda)_{\bar I} x_I\rangle|. \qquad (6.19)\end{aligned}$$

In the following, we derive bounds for the terms involved. In order to maintain a good overview we divide the proof into three steps. In the first, we derive lower and upper bounds for the first terms in (6.18) and (6.19), respectively. As a second step we bound the inner products with the perturbation dictionary Z and as a third step, put all these pieces together.

Step 1 - bounds for inner products with Ψ

For $\bar J^c = \bar G^c \cup \bar R$, for the first term in (6.18) and (6.19), we can write for any k

$$|\langle \psi_{p(k)}, Q(\Psi_{\bar J})(\Psi\Gamma)_{\bar J^c} x_{J^c}\rangle| = |\langle \psi_{p(k)}, Q(\Psi_{\bar J})(\Psi\Gamma)_{\bar G^c} x_{G^c}\rangle + \langle \psi_{p(k)}, Q(\Psi_{\bar J})(\Psi\Gamma)_{\bar R} x_R\rangle|.$$

For any index ℓ, we define $G^c_\ell = G^c \setminus \{\ell\}$, $R_\ell = R \setminus \{\ell\}$ and $\bar G^c_\ell = p(G^c_\ell)$, $\bar R_\ell = p(R_\ell)$. Note that, for $\ell \notin G^c$ we have $G^c_\ell = G^c$ and similarly for $\ell \notin R$ we have $R_\ell = R$. Using these definitions, we get for i

$$\begin{aligned}&|\langle \psi_{p(i)}, Q(\Psi_{\bar J})(\Psi\Gamma)_{\bar J^c} x_{J^c}\rangle| \\ &\quad = |\langle \psi_{p(i)}, Q(\Psi_{\bar J})\psi_{p(i)}\rangle \gamma_{p(i)} x_i + \langle \psi_{p(i)}, Q(\Psi_{\bar J})(\Psi\Gamma)_{\bar G^c_i} x_{G^c_i}\rangle \\ &\qquad + \langle \psi_{p(i)}, (\Psi\Gamma)_{\bar R} x_R\rangle - \langle \psi_{p(i)}, P(\Psi_{\bar J})(\Psi\Gamma)_{\bar R} x_R\rangle| \\ &\quad \geq \gamma_{p(i)} c_i \|Q(\Psi_{\bar J})\psi_{p(i)}\|_2^2 - |\langle \psi_{p(i)}, Q(\Psi_{\bar J})(\Psi\Gamma)_{\bar G^c_i} x_{G^c_i}\rangle| \\ &\qquad - |\langle \psi_{p(i)}, (\Psi\Gamma)_{\bar R} x_R\rangle| - |\langle \psi_{p(i)}, P(\Psi_{\bar J})(\Psi\Gamma)_{\bar R} x_R\rangle|. \qquad (6.20)\end{aligned}$$

Similarly, we have for any $k \in R$

$$\begin{aligned}&|\langle \psi_{p(k)}, Q(\Psi_{\bar J})(\Psi\Gamma)_{\bar J^c} x_{J^c}\rangle| \\ &\quad = |\langle \psi_{p(k)}, Q(\Psi_{\bar J})(\Psi\Gamma)_{\bar G^c} x_{G^c}\rangle + \gamma_{p(k)} x_k \\ &\qquad + \langle \psi_{p(k)}, (\Psi\Gamma)_{\bar R_k} x_{R_k}\rangle - \langle \psi_{p(k)}, P(\Psi_{\bar J})(\Psi\Gamma)_{\bar R} x_R\rangle| \\ &\quad \leq \gamma_{p(k)} c_k + |\langle \psi_{p(k)}, Q(\Psi_{\bar J})(\Psi\Gamma)_{\bar G^c} x_{G^c}\rangle| \\ &\qquad + |\langle \psi_{p(k)}, (\Psi\Gamma)_{\bar R_k} x_{R_k}\rangle| + |\langle \psi_{p(k)}, P(\Psi_{\bar J})(\Psi\Gamma)_{\bar R} x_R\rangle|. \qquad (6.21)\end{aligned}$$

We now bound all the terms in (6.20) and (6.21).

For the first norm term on the right-hand side of (6.20), we get

$$\begin{aligned}\|Q(\Psi_{\bar J})\psi_{p(i)}\|_2^2 &= 1 - \|P(\Psi_{\bar J})\psi_{p(i)}\|_2^2 \\ &= 1 - \langle \psi_{p(i)}, \Psi_{\bar J}(\Psi_{\bar J}^\star \Psi_{\bar J})^{-1}\Psi_{\bar J}^\star \psi_{p(i)}\rangle \\ &= 1 - \langle \Psi_{\bar J}^\star \psi_{p(i)}, (\Psi_{\bar J}^\star \Psi_{\bar J})^{-1}\Psi_{\bar J}^\star \psi_{p(i)}\rangle \\ &\geq 1 - \|\Psi_{\bar J}^\star \psi_{p(i)}\|_2 \cdot \|(\Psi_{\bar J}^\star \Psi_{\bar J})^{-1}\|_{2,2} \cdot \|\Psi_{\bar J}^\star \psi_{p(i)}\|_2. \qquad (6.22)\end{aligned}$$

Defining $\delta_{\bar{J}} := \delta_{\bar{J}}(\Psi) = \|\Psi_{\bar{J}}^\star \Psi_{\bar{J}} - \mathbb{I}\|_{2,2}$, we have by Lemma 6.2 in [24]

$$\|(\Psi_{\bar{J}}^\star \Psi_{\bar{J}})^{-1}\|_{2,2} \leq \frac{1}{1-\delta_{\bar{J}}}, \tag{6.23}$$

and therefore,

$$\|Q(\Psi_{\bar{J}})\psi_{p(i)}\|_2^2 \geq 1 - \frac{1}{1-\delta_{\bar{J}}} \|\Psi_{\bar{J}}^\star \psi_{p(i)}\|_2^2. \tag{6.24}$$

In order to bound the remaining norm term in (6.24), we use some probabilistic bound. Note that, for $\bar{J} \subseteq \bar{I}$ and any $p(j) \notin \bar{J}$, we have

$$\begin{aligned} \|\Psi_{\bar{J}}^\star \psi_{p(j)}\|_2^2 &= \sum_{k \in \bar{J}} \left|\langle \psi_k, \psi_{p(j)} \rangle\right|^2 \leq \max_{p(j)} \sum_{\substack{k \in \bar{J} \\ k \neq p(j)}} \left|\langle \psi_k, \psi_{p(j)} \rangle\right|^2 \\ &\leq \max_{p(j)} \sum_{\substack{k \in \bar{I} \\ k \neq p(j)}} \left|\langle \psi_k, \psi_{p(j)} \rangle\right|^2 = \|(\Psi^\star\Psi - \mathbb{I})_{\bar{I}}\|_{1,2}^2. \end{aligned} \tag{6.25}$$

Using Proposition 6.2 we now show that with high probability this term is small. In particular, for j fixed we choose $v = (\Psi^\star\Psi - \mathbb{I})_j = \Psi^\star\psi_j - e_j$ and use the bounds

$$\|v\|_\infty^2 \leq \bar{\mu}^2 \qquad \text{and} \qquad \|v\|_4^4 \leq \|v\|_\infty^2 \cdot \|v\|_2^2 \leq \bar{\mu}^2 \cdot \|\Psi^\star\psi_j\|_2^2 \leq \bar{\mu}^2 \bar{B}.$$

Hence, from Proposition 6.2 we obtain

$$\begin{aligned} \mathbb{P}\left(\|v_{\bar{I}}\|_2^2 \geq \frac{S\bar{B}}{K} + t\right) &\leq \exp\left(-\frac{t^2}{2\big(t\bar{\mu}^2 + \bar{\mu}^2 \frac{S\bar{B}}{K}\big)}\right) \\ &\leq \exp\left(-\frac{1}{4} \cdot \min\left\{\frac{t}{\bar{\mu}^2}, \frac{t^2 \cdot K}{\bar{\mu}^2 S\bar{B}}\right\}\right). \end{aligned} \tag{6.26}$$

Remembering that $\|\cdot\|_{1,2}^2$ is the largest squared 2-norm of a column, we get via a union bound

$$\mathbb{P}\left(\|(\Psi^\star\Psi - \mathbb{I})_{\bar{I}}\|_{1,2}^2 \geq \frac{S\bar{B}}{K} + t\right) \leq K \exp\left(-\frac{1}{4} \cdot \min\left\{\frac{t}{\bar{\mu}^2}, \frac{t^2 \cdot K}{\bar{\mu}^2 S\bar{B}}\right\}\right), \tag{6.27}$$

and therefore, for any $\bar{J} \subseteq \bar{I}$ and $p(j) \notin \bar{J}$, we have

$$\begin{aligned} \|\Psi_{\bar{J}}^\star \psi_{p(j)}\|_2 &\leq \left(\frac{S\bar{B}}{K} + 2\bar{\mu}\sqrt{n_1 \log K} \cdot \max\left\{2\bar{\mu}\sqrt{n_1 \log K}, \sqrt{\frac{S\bar{B}}{K}}\right\}\right)^{\frac{1}{2}} \\ &\leq \left(\frac{S\bar{B}}{K} + \max\left\{4\bar{\mu}^2 \cdot n_1 \log K, \frac{S\bar{B}}{K}\right\}\right)^{\frac{1}{2}} \\ &\leq \max\left\{2\bar{\mu} \cdot \sqrt{2n_1 \log K}, \sqrt{\frac{2S\bar{B}}{K}}\right\} =: E, \end{aligned} \tag{6.28}$$

except with probability K^{1-n_1}. Note that, instead of using this probabilistic bound we could also use the deterministic bound $\|\Psi^\star_{\bar{J}}\psi_{p(j)}\|_2 \leq \sqrt{|J|}\bar{\mu}$. Especially for partial support recovery conditions where $|\bar{J}| = |J| \ll |I| = S$, we might get better estimates when using the crude bound $\sqrt{|J|}\bar{\mu}$. However, in case of results covering also full support recovery, the number of elements contained in $\bar{J}$ resp. J can only be bounded by $|J| \leq S$ and in general $\bar{B}/K \ll \bar{\mu}^2$.

For the sake of convenience, as it is easier to handle when combining all the estimates at the end of the proof, for the squared norm term we use the bound

$$\|\Psi^\star_{\bar{J}}\psi_{p(j)}\|_2^2 \leq \sqrt{|J|}\bar{\mu} \cdot E. \tag{6.29}$$

Hence, putting all these pieces together, for the first norm term on the right-hand side of (6.20), we finally get for i

$$\|Q(\Psi_{\bar{J}})\psi_{p(i)}\|_2^2 \geq 1 - \frac{\sqrt{|J|}\bar{\mu}}{1-\delta_{\bar{J}}} \cdot E. \tag{6.30}$$

Next, we bound the second term on the right-hand side of (6.20) and (6.21), respectively. Since the number of elements contained in the set G^c is small, we use only a crude bound for this term. In particular, for all $k \in G^c \cup R$, we have

$$\begin{aligned}
&|\langle \psi_{p(k)}, Q(\Psi_{\bar{J}})(\Psi\Gamma)_{\bar{G}^c_k} x_{G^c_k}\rangle| \\
&\quad = |\langle \psi_{p(k)}, (\Psi\Gamma)_{\bar{G}^c_k} x_{G^c_k}\rangle - \langle \Psi^\star_{\bar{J}}\psi_{p(k)}, (\Psi^\star_{\bar{J}}\Psi_{\bar{J}})^{-1}\Psi^\star_{\bar{J}}(\Psi\Gamma)_{\bar{G}^c_k} x_{G^c_k}\rangle| \\
&\quad \leq |\langle \psi_{p(k)}, (\Psi\Gamma)_{\bar{G}^c_k} x_{G^c_k}\rangle| + \|\Psi^\star_{\bar{J}}\psi_{p(k)}\|_2 \cdot \|(\Psi^\star_{\bar{J}}\Psi_{\bar{J}})^{-1}\|_{2,2} \cdot \|\Psi^\star_{\bar{J}}(\Psi\Gamma)_{\bar{G}^c_k} x_{G^c_k}\|_2 \\
&\quad \leq |\langle \psi_{p(k)}, (\Psi\Gamma)_{\bar{G}^c_k} x_{G^c_k}\rangle| + \|\Psi^\star_{\bar{J}}\psi_{p(k)}\|_2 \cdot \frac{\sqrt{|J|}}{1-\delta_{\bar{J}}} \cdot \max_{\ell \in \bar{J}} |\langle \psi_\ell, (\Psi\Gamma)_{\bar{G}^c_k} x_{G^c_k}\rangle| \\
&\quad \leq \max_{k \in G^c \cup \bar{R}} |\langle \psi_{p(k)}, (\Psi\Gamma)_{\bar{G}^c_k} x_{G^c_k}\rangle| + E \cdot \frac{\sqrt{|J|}}{1-\delta_{\bar{J}}} \cdot \max_{\ell \in \bar{J}} |\langle \psi_\ell, (\Psi\Gamma)_{\bar{G}^c_k} x_{G^c_k}\rangle| \\
&\quad \leq \bar{\mu} \cdot \|\Gamma_{\bar{G}^c_k} x_{G^c_k}\|_1 \left(1 + \frac{\sqrt{|J|}}{1-\delta_{\bar{J}}} \cdot E\right).
\end{aligned} \tag{6.31}$$

For the terms involving $P(\Psi_{\bar{J}})$, we have for any $k \in G^c \cup R$

$$\begin{aligned}
|\langle \psi_{p(k)}, P(\Psi_{\bar{J}})(\Psi\Gamma)_{\bar{R}} x_R\rangle| &= |\langle \Psi^\star_{\bar{J}}\psi_{p(k)}, (\Psi^\star_{\bar{J}}\Psi_{\bar{J}})^{-1}\Psi^\star_{\bar{J}}(\Psi\Gamma)_{\bar{R}} x_R\rangle| \\
&\leq \|\Psi^\star_{\bar{J}}\psi_{p(k)}\|_2 \cdot \|(\Psi^\star_{\bar{J}}\Psi_{\bar{J}})^{-1}\|_{2,2} \cdot \|\Psi^\star_{\bar{J}}(\Psi\Gamma)_{\bar{R}} x_R\|_2 \\
&\leq \max_{\ell \in \bar{J}} |\langle \psi_\ell, (\Psi\Gamma)_{\bar{R}} x_R\rangle| \cdot \frac{\sqrt{|J|}}{1-\delta_{\bar{J}}} \cdot \|\Psi^\star_{\bar{J}}\psi_{p(k)}\|_2 \\
&\leq \max_{\ell \notin R} |\langle \psi_{p(\ell)}, (\Psi\Gamma)_{\bar{R}} x_R\rangle| \cdot \frac{\sqrt{|J|}}{1-\delta_{\bar{J}}} \cdot E.
\end{aligned} \tag{6.32}$$

In order to bound $|\langle \psi_{p(\ell)}, (\Psi\Gamma)_{\bar{R}} x_R\rangle|$ for $\ell \notin R$ and $|\langle \psi_{p(k)}, (\Psi\Gamma)_{\bar{R}_k} x_{R_k}\rangle|$ for $k \in R$ in (6.21), we use Hoeffding's inequality. Note again that for $k \notin R$, we have $R_k = R$ and hence, we get for any k

$$\begin{aligned}
\mathbb{P}\big(|\langle \psi_{p(k)}, (\Psi\Gamma)_{\bar{R}_k} x_{R_k}\rangle| \geq \theta_k\big) &= \mathbb{P}\big(\big|\sum_{j\in R_k} \langle \psi_{p(k)}, \psi_{p(j)}\rangle \gamma_{p(j)} c_j \sigma_j\big| \geq \theta_k\big) \\
&\leq 2\exp\left(-\frac{\theta_k^2}{2\sum_{j\in R_k} \left|\langle \psi_{p(k)}, \psi_{p(j)}\rangle\right|^2 \gamma_{p(j)}^2 c_j^2}\right) \\
&\leq 2\exp\left(-\frac{\theta_k^2}{2\bar{\mu}^2 \|\Gamma_{\bar{R}} c_R\|_2^2}\right). \qquad (6.33)
\end{aligned}$$

Setting $\theta_k = 2\bar{\mu}\sqrt{n_2 \log K} \cdot \|\Gamma_{\bar{R}} c_R\|_2$ and using a union bound over all indices k, we have

$$\max_k |\langle \psi_{p(k)}, (\Psi\Gamma)_{\bar{R}_k} x_{R_k}\rangle| \leq 2\bar{\mu}\sqrt{n_2 \log K} \cdot \|\Gamma_{\bar{R}} c_R\|_2, \qquad (6.34)$$

except with probability $2K^{1-2n_2}$.

Note that, this bound holds only for one specific $\bar{R} := p(R) := p(R_i)$ however, in order to show that OMP succeeds we need this bound for all possible R. As R is defined by A_i, for this inequality, we will use another union bound over all sets A_i at the very end of the proof.

Combining these estimates, for the last two terms in (6.20) and (6.21), respectively, we have for any $k \in G^c \cup R$,

$$\begin{aligned}
&|\langle \psi_{p(k)}, (\Psi\Gamma)_{\bar{R}_k} x_{R_k}\rangle| + |\langle \psi_{p(k)}, P(\Psi_{\bar{J}})(\Psi\Gamma)_{\bar{R}} x_R\rangle| \\
&\qquad\qquad \leq 2\bar{\mu}\sqrt{n_2 \log K} \cdot \|\Gamma_{\bar{R}} c_R\|_2 \left(1 + \frac{\sqrt{|J|}}{1-\delta_{\bar{J}}} \cdot E\right), \qquad (6.35)
\end{aligned}$$

except with probability $2K^{1-2n_2}$.

Putting the bounds from (6.30), (6.31) and (6.35) together, for the first part in (6.18) resp. (6.19), we get for i

$$\begin{aligned}
&|\langle \psi_{p(i)}, Q(\Psi_{\bar{J}})(\Psi\Gamma)_{\bar{J}^c} x_{J^c}\rangle| \\
&\quad \geq \gamma_{p(i)} c_i \left(1 - \frac{\sqrt{|J|}\bar{\mu}}{1-\delta_{\bar{J}}} \cdot E\right) \\
&\qquad\quad - \left(\|\Gamma_{\bar{G}_i^c} x_{G_i^c}\|_1 + 2\sqrt{n_2 \log K} \cdot \|\Gamma_{\bar{R}} c_R\|_2\right)\left(\bar{\mu} + \frac{\sqrt{|J|}\bar{\mu}}{1-\delta_{\bar{J}}} \cdot E\right) \\
&\quad \geq \gamma_{p(i)} c_i - \left(\|\Gamma_{\bar{G}^c} x_{G^c}\|_1 + 2\sqrt{n_2 \log K} \cdot \|\Gamma_{\bar{R}} c_R\|_2\right)\left(\bar{\mu} + \frac{\sqrt{|J|}\bar{\mu}}{1-\delta_{\bar{J}}} \cdot E\right), \qquad (6.36)
\end{aligned}$$

and all $k \in R$,

$$
\begin{aligned}
&|\langle \psi_{p(k)}, Q(\Psi_{\bar{J}})(\Psi\Gamma)_{\bar{J}^c} x_{J^c}\rangle| \\
&\qquad \le \gamma_{p(k)} c_k + \Big(\|\Gamma_{\bar{G}^c} x_{G^c}\|_1 + 2\sqrt{n_2 \log K} \cdot \|\Gamma_{\bar{R}} c_R\|_2 \Big) \left(\bar{\mu} + \frac{\sqrt{|J|}\bar{\mu}}{1-\delta_{\bar{J}}} \cdot E \right), \qquad (6.37)
\end{aligned}
$$

except with probability $2K^{1-2n_2}$.

As a next step we derive bounds for the remaining terms in (6.18) and (6.19), meaning for the inner products with the perturbation dictionary Z.

Step 2 - bounds for inner products with Z

For the second term in (6.18) and (6.19), for all $k \in G^c \cup R$, we get

$$
\begin{aligned}
\big|\langle \psi_{p(k)}, Q(\Psi_{\bar{J}})(Z\Lambda)_{\bar{I}} x_I\rangle\big| &= |\langle \psi_{p(k)}, (Z\Lambda)_{\bar{I}} x_I\rangle - \langle \psi_{p(k)}, P(\Psi_{\bar{J}})(Z\Lambda)_{\bar{I}} x_I\rangle| \\
&\le |\langle \psi_{p(k)}, (Z\Lambda)_{\bar{I}} x_I\rangle| + |\langle \Psi_{\bar{J}}^\star \psi_{p(k)}, (\Psi_{\bar{J}}^\star \Psi_{\bar{J}})^{-1} \Psi_{\bar{J}}^\star (Z\Lambda)_{\bar{I}} x_I\rangle| \\
&\le |\langle \psi_{p(k)}, (Z\Lambda)_{\bar{I}} x_I\rangle| + \frac{\sqrt{|J|}}{1-\delta_{\bar{J}}} \cdot \|\Psi_{\bar{J}}^\star \psi_{p(k)}\|_2 \cdot \max_{\ell \in \bar{J}} |\langle \psi_\ell, (Z\Lambda)_{\bar{I}} x_I\rangle| \\
&\le \max_{\ell} |\langle \psi_{p(\ell)}, (Z\Lambda)_{\bar{I}} x_I\rangle| \cdot \left(1 + \frac{\sqrt{|J|}}{1-\delta_{\bar{J}}} \cdot E \right). \qquad (6.38)
\end{aligned}
$$

In order to bound the inner product in the above inequality, we use Proposition 3.4. Note that, we have $I = \{1, \ldots, S\}$ and hence, for any index ℓ

$$
|\langle \psi_{p(\ell)}, (Z\Lambda)_{\bar{I}} x_I\rangle| = \big| \sum_{k \in I} \langle \psi_{p(\ell)}, z_{p(k)}\rangle \lambda_{p(k)} c_k \sigma_k \big| = \big| \sum_{k=1}^{S} \langle \psi_{p(\ell)}, z_{p(k)}\rangle \lambda_{p(k)} c_k \sigma_k \big|. \qquad (6.39)
$$

Defining $\bar{\nu}_Z = \max_{i,j} |\langle \psi_i, z_j \lambda_j\rangle|$, and setting $v_{i_k} = v_{p(k)} = \langle \psi_{p(\ell)}, z_{p(k)}\rangle \lambda_{p(k)}$ and $c_k \varepsilon_k = c_k \sigma_k$, from Proposition 3.4, we get for any index ℓ

$$
\begin{aligned}
&\mathbb{P}\big(\big| \sum_{k=1}^{S} \langle \psi_{p(\ell)}, z_{p(k)}\rangle \lambda_{p(k)} c_k \sigma_k \big| \ge t \big) \\
&\qquad\qquad \le 2 \exp\left(- \frac{t^2}{2\Big(\|c\|_\infty \|\psi_{p(\ell)}^\star Z\Lambda\|_\infty \cdot t + \|c\|_2^2 \cdot \frac{\|\psi_{p(\ell)}^\star Z\Lambda\|_2^2}{K-S} \Big)} \right) \\
&\qquad\qquad \le 2 \exp\left(- \frac{t^2}{2\Big(\|c\|_\infty \cdot \bar{\nu}_Z \cdot t + \|c\|_2^2 \cdot \frac{\|Z\Lambda\|_{2,2}^2}{K-S} \Big)} \right) \\
&\qquad\qquad \le 2 \exp\left(-\frac{1}{4} \cdot \min\left\{ \frac{t}{\|c\|_\infty \cdot \bar{\nu}_Z}, \frac{t^2 (K-S)}{\|c\|_2^2 \cdot \|Z\Lambda\|_{2,2}^2} \right\} \right). \qquad (6.40)
\end{aligned}
$$

Together with a union bound over all indices ℓ and using $\|c\|_2 = \|c_I\|_2$, we have

$$\max_{\ell} |\langle \psi_{p(\ell)}, (Z\Lambda)_{\bar{I}} x_I \rangle| \leq 2\sqrt{n_3 \log K} \cdot \max \left\{ 2\bar{\nu}_Z \|c_I\|_\infty \sqrt{n_3 \log K}, \|c_I\|_2 \frac{\|Z\Lambda\|_{2,2}}{\sqrt{K-S}} \right\}.$$

except with probability $2K^{1-n_3}$.

Combining these estimates, for the inner products of the perturbation part, we have for any index k

$$\begin{aligned} &|\langle \psi_{p(k)}, Q(\Psi_{\bar{J}})(Z\Lambda)_{\bar{I}} x_I \rangle| \\ &\quad \leq 2\sqrt{n_3 \log K} \cdot \max \left\{ 2\bar{\nu}_Z \|c_I\|_\infty \sqrt{n_3 \log K}, \|c_I\|_2 \frac{\|Z\Lambda\|_{2,2}}{\sqrt{K-S}} \right\} \left(1 + \frac{E\sqrt{|J|}}{1-\delta_{\bar{J}}} \right), \end{aligned}$$

except with probability $2K^{1-n_3}$.

Step 3 - combining all the estimates

In order to get a sufficient condition ensuring that OMP adds another correct atom within the subsequent iteration, we now have to combine all the previously obtained estimates. Hence, for i and any $k \in R$, we have

$$\begin{aligned} &\big|\langle \psi_{p(i)}, r_{\bar{J}} \rangle\big| - \big|\langle \psi_{p(k)}, r_{\bar{J}} \rangle\big| \\ &\geq \gamma_{p(i)} c_i - \gamma_{p(k)} c_k \\ &\quad - 4\sqrt{n_3 \log K} \cdot \max \left\{ 2\bar{\nu}_Z \|c_I\|_\infty \sqrt{n_3 \log K}, \|c_I\|_2 \frac{\|Z\Lambda\|_{2,2}}{\sqrt{K-S}} \right\} \left(1 + \frac{E\sqrt{|J|}}{1-\delta_{\bar{J}}} \right) \\ &\qquad - 2\bar{\mu} \cdot \left(\|\Gamma_{\bar{G}^c} c_{G^c}\|_1 + 2\sqrt{n_2 \log K} \cdot \|\Gamma_{\bar{R}} c_R\|_2 \right) \left(1 + \frac{E\sqrt{|J|}}{1-\delta_{\bar{J}}} \right), \end{aligned} \tag{6.41}$$

except with probability $2K(K^{-2n_2} + K^{-n_3})$.

To get a bound for the constant $\delta_{\bar{J}} \leq \|\Psi_{\bar{I}}^\star \Psi_{\bar{I}} - \mathbb{I}\|_{2,2} = \delta(\Psi_{\bar{I}}) =: \delta_{\bar{I}}$, we use Chretien and Darses's result on the conditioning of random subdictionaries. In particular, using Theorem 3.1 of [14], reformulated for our purposes, we have

$$\begin{aligned} \mathbb{P}\left(\delta(\Psi_{\bar{I}}) > \delta_0 \,\big|\, |\bar{I}| = S \right) &\leq 216K \cdot \exp\left(-\min\left\{ \frac{\delta_0}{2\bar{\mu}}, \frac{\delta_0^2 K}{4e^2 S \bar{B}} \right\} \right) \\ &\leq 216K \cdot \max\{K^{-n_4}, K^{-n_5}\}, \end{aligned}$$

whenever,

$$\bar{\mu} \leq \frac{\delta_0}{2n_4 \log K} \quad \text{and} \quad S \leq \frac{\delta_0^2 K}{4n_5 e^2 \bar{B} \log K}.$$

Note that, for the term E in (6.28), we have $E \leq \sqrt{2S\bar{B}/K}$ whenever $4\bar{\mu}^2 n_1 \log K \leq S\bar{B}/K$. Defining $\gamma_{\min} = \min_k \gamma_k$, $\varepsilon_{\min} = \min_k \varepsilon_k$, $\gamma = \max_k \gamma_k$ and $\varepsilon = \max_k \varepsilon_k$, we have for $\varepsilon \leq 1$,

$$\gamma_j = \frac{2}{2-\varepsilon_j^2} \geq \frac{2}{2-\varepsilon_{\min}^2} = \gamma_{\min} \geq 1, \tag{6.42}$$

$$\gamma_k = \frac{2}{2-\varepsilon_k^2} \leq \frac{2}{2-\varepsilon^2} = \gamma \leq 2, \tag{6.43}$$

$$\lambda_k = \gamma_k \left(\varepsilon_k^2 - \frac{\varepsilon_k^4}{4}\right)^{\frac{1}{2}} \leq \gamma_k \varepsilon_k \leq \gamma\varepsilon, \tag{6.44}$$

and hence, for $\nu_Z = \max_{i,j} |\langle \psi_i, z_j \rangle|$,

$$\bar{\nu}_Z = \max_{i,j} |\langle \psi_i, z_j \lambda_j \rangle| \leq \gamma\varepsilon \cdot \nu_Z \qquad \text{and} \qquad \|Z\Lambda\|_{2,2} \leq \gamma\varepsilon \cdot \|Z\|_{2,2}. \tag{6.45}$$

Setting $n_1 = n_2 = n_3 = n$ and $n_4 = n_5 = m$, for $\delta_{\bar{J}} \leq \delta_{\bar{I}} \leq \frac{1}{2}$ and $4\bar{\mu}^2 n \log K \leq S\bar{B}/K$, we can further bound

$$\begin{aligned}
&\left|\langle \psi_{p(i)}, r_{\bar{J}} \rangle\right| - \left|\langle \psi_{p(k)}, r_{\bar{J}} \rangle\right| \\
&\geq c_i \gamma_{\min} - c_k \gamma \\
&\quad - 4\gamma\varepsilon\sqrt{n \log K} \cdot \max\left\{2\nu_Z \|c_I\|_\infty \sqrt{n \log K}, \|c_I\|_2 \frac{\|Z\|_{2,2}}{\sqrt{K-S}}\right\}\left(1 + 2\sqrt{\frac{2|J|S\bar{B}}{K}}\right) \\
&\qquad - 2\bar{\mu} \cdot \left(\|\Gamma_{\bar{G}^c} c_{G^c}\|_1 + 2\sqrt{n \log K} \cdot \|\Gamma_{\bar{R}} c_R\|_2\right)\left(1 + 2\sqrt{\frac{2|J|S\bar{B}}{K}}\right),
\end{aligned} \tag{6.46}$$

except with probability $K(2K^{-2n} + 3K^{-n} + 216K^{-m})$. To bound the norm terms of the coefficients, we use the definitions of the slots introduced in Section 6.1.1, where we have

$$b(i) = j \Leftrightarrow c_i \in (c_1\beta^j, c_1\beta^{j-1}] \quad \text{and} \quad U_j = b^{-1}(\{j\}).$$

Hence, for $G^c = C^c \cup D^c$, $t := \max_j |U_j|$ the maximal number of elements in a slot, and $c_i \in C^c$ the largest still missing coefficient, we get

$$\begin{aligned}
\|\Gamma_{\bar{G}^c} c_{G^c}\|_1 &= \sum_{\ell \in G^c} |\gamma_{p(\ell)} c_\ell| \leq \gamma \left(\sum_{\ell \in C^c} |c_\ell| + \sum_{\ell \in D^c} |c_\ell|\right) \\
&\leq \gamma\left(c_i |C^c| + c_1 \beta^{b(i)} |D^c|\right) \leq c_i \cdot \gamma\left(|C^c| + |D^c|\right) \leq c_i \cdot \gamma \cdot 2t.
\end{aligned} \tag{6.47}$$

For all $k \in R$, we have $c_k \leq c_1\beta^{b(i)+1}$ and therefore,

$$\begin{aligned}\|\Gamma_{\bar{R}} c_R\|_2 &\leq c_1\beta^{b(i)+1} \cdot \gamma \left(t \cdot \sum_{k=0}^{\infty} \beta^{2k} \right)^{\frac{1}{2}} \leq c_1\beta^{b(i)+1} \cdot \gamma \left(\frac{t}{1-\beta^2} \right)^{\frac{1}{2}} \\ &\leq c_i\beta \cdot \gamma \left(\frac{t}{1-\beta^2} \right)^{\frac{1}{2}}. \end{aligned} \tag{6.48}$$

Inserting these bounds into (6.46) and using that $c_1\beta^{b(i)} \leq c_i$ and hence, $\frac{\|c_I\|_\infty}{c_i} \leq \frac{\|c_I\|_\infty}{c_1\beta^{b(i)}}$ and $\frac{\|c_I\|_2}{c_i} \leq \frac{\|c_I\|_2}{c_1\beta^{b(i)}}$, we obtain

$$\begin{aligned} & c_i^{-1} \left(\left|\langle \psi_{p(i)}, r_{\bar{J}} \rangle\right| - \left|\langle \psi_{p(k)}, r_{\bar{J}} \rangle\right| \right) \\ & \geq \gamma_{\min} - \gamma\beta \\ & \quad - 4\gamma\varepsilon\sqrt{n\log K} \cdot \max\left\{ 2\nu_Z \frac{\|c_I\|_\infty}{c_1\beta^{b(i)}} \sqrt{n\log K}, \frac{\|c_I\|_2}{c_1\beta^{b(i)}} \frac{\|Z\|_{2,2}}{\sqrt{K-S}} \right\} \left(1 + 2\sqrt{\frac{2|J|S\bar{B}}{K}} \right) \\ & \qquad - 2\gamma \cdot \bar{\mu} \cdot \left(2t + 2\sqrt{n\log K} \cdot \left(\frac{t\beta^2}{1-\beta^2} \right)^{\frac{1}{2}} \right) \left(1 + 2\sqrt{\frac{2|J|S\bar{B}}{K}} \right), \end{aligned} \tag{6.49}$$

except with probability $K(2K^{-2n} + 3K^{-n} + 216K^{-m})$.

Using that $|J| \leq S$, in case $4\bar{\mu}^2 n\log K \leq S\bar{B}/K$, a condition ensuring that with high probability OMP picks $p(i)$ before $p(k)$ for any $k \in R$, is therefore given by

$$\begin{aligned} & \gamma_{\min} - \gamma\beta \\ & \overset{!}{\geq} 4\gamma\varepsilon\sqrt{n\log K} \cdot \max\left\{ 2\nu_Z \frac{c_1}{c_1\beta^{b(i)}} \sqrt{n\log K}, \frac{\|c_I\|_2}{c_1\beta^{b(i)}} \frac{\|Z\|_{2,2}}{\sqrt{K-S}} \right\} \left(1 + 2S\sqrt{\frac{2\bar{B}}{K}} \right) \\ & \qquad + 4\gamma \cdot \bar{\mu} \cdot \left(t + \sqrt{n\log K} \cdot \left(\frac{t\beta^2}{1-\beta^2} \right)^{\frac{1}{2}} \right) \left(1 + 2S\sqrt{\frac{2\bar{B}}{K}} \right). \end{aligned} \tag{6.50}$$

If we look at the next step, with i_{new} the index of the new largest missing coefficient, we can reuse the bounds from above, except for (6.34), as they do not depend on the decomposition, meaning, on i_{new}. However, the bound in (6.34) depends on i_{new} and so we need this for all decompositions. Hence, taking a union bound over all possible sets A_i and thus all possible sets $R := R_i$ in (6.34), choosing $c_1\beta^{b(i)} = c_1\beta^s$ and multiplying both sides by $\gamma_{\min}^{-1}$ finally yields the result. The condition in case where $4\bar{\mu}^2 n\log K \geq S\bar{B}/K$ is obtained by replacing $2S\sqrt{\frac{2\bar{B}}{K}}$ with $4\bar{\mu}\sqrt{2Sn\log K}$. □

As a next step we derive recovery conditions for OMP using a perturbed dictionary for the case of noisy signals.

6.3 Recovery Conditions for Noisy Signals

Here we derive conditions ensuring partial support recovery in case of signals contaminated with noise. For that, let our signals be modelled as

$$\tilde{y} = y + \eta = \Phi_{\bar{I}} x_I + \eta = \sum_{i \in I} \phi_{p(i)} \sigma_i c_i + \eta, \tag{6.51}$$

with y defined as in (6.1) and η a sub-Gaussian noise vector with parameter ρ. In particular, this means that we have $\mathbb{E}(\eta) = 0$ and for all vectors v with $\|v\|_2 = 1$ and $\theta > 0$ the marginals $\langle v, \eta \rangle$ satisfy $\mathbb{E}(e^{\theta \langle v, \eta \rangle}) \leq e^{\theta^2 \rho^2 / 2}$.
In case of Gaussian noise, the parameter ρ corresponds to the standard deviation and hence, for normalised coefficient sequences c, $\|c\|_2 = 1$, the signal to noise ratio (SNR) is $\frac{1}{d\rho^2}$. Note that, with $\rho = \sqrt{B_u} \|c_{I^c}\|_\infty$, where B_u denotes the upper frame bound of Φ and $I^c = \{S+1, \dots, K\}$, the signal model in (6.51) provides also a generalisation of (6.1) to approximately S-sparse signals, where $y = \sum_i \phi_{p(i)} \sigma_i c_i$.

Theorem 6.3. *Assume that the signals follow the model in (6.51) with coefficients grouped by their magnitude into s slots U_j as defined in (6.4). Let t denote the maximal number of elements within a slot and Ψ some perturbation of the generating dictionary Φ as defined in (6.7) with $\varepsilon \leq 1$. Further, assume that*

$$\mu(\Psi) \leq \frac{1}{4m \log K} \qquad \textit{and} \qquad S \leq \frac{K}{16 m e^2 \|\Psi\|_{2,2}^2 \log K}.$$

Then OMP will recover an atom from the support in the first ℓ iterations, except with probability $K(4sK^{-2n} + 2K^{-n} + 216K^{-m})$, as long as

$$\begin{aligned}
1 - \frac{\gamma}{\gamma_{min}} \beta & \\
\geq 4\, \varepsilon \cdot \frac{\gamma}{\gamma_{min}} \sqrt{n \log K} & \cdot \max \left\{ 2\nu_Z \frac{\|c_I\|_\infty}{c_\ell} \sqrt{n \log K}, \frac{\|c_I\|_2}{c_\ell} \sqrt{\frac{\|Z\|_{2,2}^2}{K-S}} \right\} (1 + 2\ell\mu(\Psi)) \\
& + 4\mu(\Psi) \cdot \frac{\gamma}{\gamma_{min}} \left(t + \sqrt{nt \log K} \cdot \left(\frac{\beta^2}{1-\beta^2} \right)^{\frac{1}{2}} \right) (1 + 2\ell\mu(\Psi)) \\
& + 4 \frac{\rho}{\gamma_{min} \cdot c_\ell} \sqrt{n \log K} \left(1 + 2\sqrt{2t\ell} \cdot \mu(\Psi) \right),
\end{aligned} \tag{6.52}$$

where $\varepsilon = \max_k \|\phi_k - \psi_k\|_2$, $\gamma = \frac{2}{2-\varepsilon^2}$, $\gamma_{min} = \frac{2}{2-\varepsilon_{min}^2}$ with $\varepsilon_{min} = \min_k \|\phi_k - \psi_k\|_2$ and $\nu_Z = \max_{i,j} |\langle \psi_i, z_j \rangle|$.

Considering the case of no perturbation, meaning, where $\Psi = \Phi$, the condition ensuring partial support recovery in the first ℓ iterations, except with probability

$K(4sK^{-2n}+216K^{-m})$, is given by

$$4\mu(\Psi)\left(t+\sqrt{nt\log K}\cdot\left(\frac{\beta^2}{1-\beta^2}\right)^{\frac{1}{2}}\right)(1+2\ell\mu(\Psi)) \\ +4\frac{\rho}{c_\ell}\sqrt{n\log K}\left(1+2\sqrt{2t\ell}\cdot\mu(\Psi)\right)\leq 1-\beta. \tag{6.53}$$

From this we can see that even in the perturbation-free case, OMP is only able to recover atoms corresponding to coefficients above the noise-level. More precisely, the condition in (6.53) puts also a constraint on the smallest coefficient c_ℓ which we are able to recover as in order to have the second term sufficiently small we need to have $c_\ell \geq \kappa\rho\sqrt{n\log K}$, where κ denotes some constant. Note that, the condition for the perturbation-free case is equivalent to the condition derived in [60] however, (6.53) formulates a generalisation as it is valid for a larger class of signals.

In order not to get lost in too many details and to maintain a better overview, we want to refer to Appendix A.3 for the proof of Theorem 6.3. To get a better feeling for the terms involved, let us specialise the condition in (6.52) to the case where $t=1$ and therefore, $c_\ell > c_1\beta^\ell$. Further, assume that all atoms are equally perturbed, $\varepsilon_{\min}\approx\varepsilon$, $\|Z\|_{2,2}\lesssim\sqrt{\log K}$ and $2\nu_Z\|c_I\|_\infty \lesssim \|c_I\|_2/\sqrt{K-S}$. Hence, the condition ensuring the recovery of an atom from the support in the first ℓ iterations is approximately given by

$$\frac{1-\beta}{4}\gtrsim\sqrt{n\log K}\left(\mu(\Psi)+\varepsilon\cdot\beta^{-\ell}\frac{\|c_I\|_2}{c_1}\sqrt{\frac{\log K}{K-S}}+\beta^{-\ell}\frac{\rho}{c_1}\right)(1+2\ell\mu(\Psi)). \tag{6.54}$$

From this we can again see that strongly decaying signal coefficients and hence, β small, decrease the success rate of OMP as they increase the contribution of the perturbation and noise parts. While even in the perturbation-free case, $\varepsilon=0$, the range of parameters for which OMP performs well is limited by the noise level ρ, this limitation is further increased with increasing ε.

6.4 Comparison of OMP and Thresholding

In this section we provide recovery conditions for thresholding in case of noiseless perfectly S-sparse signals and compare them with the conditions obtained for OMP. Note that, implicitly we already have them from the proof of Theorem 3.1 in Chapter 3. However, here we restate them in a shape that is better comparable to the conditions of OMP and for completeness also provide the proof in the appendix.

Theorem 6.4. *Assume that the signals follow the model in (6.1) with $c_S=\min_{i\in I}c_i$. Then, given a perturbation Ψ of the generating dictionary Φ as defined in (6.7) and*

$\varepsilon \leq 1$, *we have that except with probability* $4K^{1-2n}$, *thresholding recovers the generating support* $\bar{I}$ *if*

$$4\frac{\gamma}{\gamma_{min}}(\mu(\Psi) + \varepsilon\nu_Z) \cdot \frac{\|c_I\|_2}{c_S}\sqrt{n\log K} \leq 1, \tag{6.55}$$

where $\varepsilon = \max_k \|\phi_k - \psi_k\|_2$, $\gamma = \frac{2}{2-\varepsilon^2}$ *and* $\gamma_{min} = \frac{2}{2-\varepsilon_{min}^2}$ *with* $\varepsilon_{min} = \min_k \|\phi_k - \psi_k\|_2$ *and* $\nu_Z = \max_{i,j} |\langle\psi_i, z_j\rangle|$.

From the condition in (6.55) we see that the success of thresholding depends on the quantity $\frac{\|c_I\|_2}{c_S}$. For equally sized coefficients we have $\frac{\|c_I\|_2}{c_S} = \sqrt{S}$, however, in more realistic scenarios where the coefficients are not of equal size this term can grow very fast. In particular, in case of decaying coefficients we in general have $c_S \ll 1$, especially for larger S. Even in the perturbation free-case, $\varepsilon = 0$, the dependence on this term strongly restricts the range of parameters for which thresholding performs well. The proof of Theorem 6.4 can be found in Appendix A.4.

A major advantage of thresholding is its low computational complexity. Thresholding is computationally much cheaper than OMP. However, when we have the signal generating dictionary, OMP is known to perform better especially for signals with coefficients which are not of equal size, [60].

In Chapter 5 we have seen that ITKrM provides a computationally much lighter alternative to K-SVD. Interestingly, although K-SVD uses the better (even though more expensive) sparse approximation algorithm OMP, both algorithms were shown to yield similar results. To get some theoretical insights which shed light on the question why K-SVD performs not much better than ITKrM we will now compare the recovery conditions of their corresponding sparse approximation algorithms.

Note that, compared to thresholding where all the S most contributing atoms are picked at once, in each iteration, OMP always adds one atom at the time. It is therefore difficult to directly compare the corresponding recovery conditions of these algorithms. However, in order to still gain a rough comparison, let us consider the situation where we assume that OMP has correctly recovered all atoms except for some with indices within the last slot U_s. More precisely, using the definition of the slots introduced in Section 6.1.1 with i the index of the largest missing coefficient, we assume that $c_i \in U_s$, $N_i = \emptyset$ and $R_i = \mathbb{S}^c$. Note that, we have $c_i \geq c_S \in U_s$ and $c_k = 0$ for all $k \in R_i$. Using that $\|Z\|_{2,2}/\sqrt{K-S} \lesssim \mu(\Psi, Z) \leq \nu_Z$ and assuming that in each slot we have no more than $\sqrt{n\log K}$ elements, in case of exactly S-sparse noiseless signals, the condition ensuring that OMP recovers c_S with high probability is approximately

$$\begin{aligned} 1 \gtrsim 4\frac{\gamma}{\gamma_{\min}} \cdot \varepsilon\nu_Z \cdot \frac{\|c_I\|_2}{c_S}\sqrt{n\log K}\,(1+2S\mu(\Psi)) \\ + 2\frac{\gamma}{\gamma_{\min}} \cdot \mu(\Psi) \cdot \sqrt{n\log K}\,(1+2S\mu(\Psi)), \end{aligned} \tag{6.56}$$

whereas for thresholding, we need to have

$$\begin{aligned} 1 \geq 4\frac{\gamma}{\gamma_{\min}} \cdot \varepsilon\nu_Z \cdot \frac{\|c_I\|_2}{c_S}\sqrt{n\log K} \\ + 4\frac{\gamma}{\gamma_{\min}} \cdot \mu(\Psi) \cdot \frac{\|c_I\|_2}{c_S}\sqrt{n\log K}. \end{aligned} \tag{6.57}$$

Comparing these conditions, we see that for $S\mu(\Psi) \leq 1$ the first parts of the right hand sides are approximately the same. In particular, with increasing ε the term $\varepsilon\nu_Z \cdot \frac{\|c_I\|_2}{c_S}$ starts to dominate both conditions. Therefore, for large ε the recovery condition of OMP is very similar to the one of thresholding. This may explain why OMP works very well in case where we have the generating dictionary ($\varepsilon = 0$) and the fast decrease of its performance if we have to deal with perturbations ($\varepsilon > 0$). In order to verify these and also our previous observations, in the following we conduct some numerical experiments.

6.5 Numerical Simulations

In order to see how some perturbation added to the generating dictionary Φ influences the performance of OMP, we conduct some numerical experiments with noiseless as well as noisy signals in $\mathbb{R}^d$, with $d = 128$. In particular, we assume that the signals follow the model in (6.1) and (6.51), respectively, where the permutation p is chosen uniformly at random. For the signal coefficients we consider the special case where they form a geometric sequence with decay factor $\alpha \in [0.75, 1]$, that is, $c_i = \kappa_S\alpha^i$ for $i \leq S$ and $c_i = 0$ for all $i > S$, where κ_S denotes some constant ensuring that $\|c\|_2 = 1$. The sparsity level S is chosen between 2 and 48 and for the noisy signals we choose η to be i.i.d. Gaussian with variance $\rho^2 = \frac{1}{256d}$ and $\rho^2 = \frac{1}{16d}$. For the generating dictionary Φ we use the concatenation of the Dirac and DCT bases as well as the Dirac-DCT dictionary with additional $2d$ vectors chosen uniformly at random from the unit sphere. The perturbed dictionary Ψ is obtained by adding some scaled perturbation dictionary Z to the generating dictionary Φ, as defined in (6.7). In particular, we have

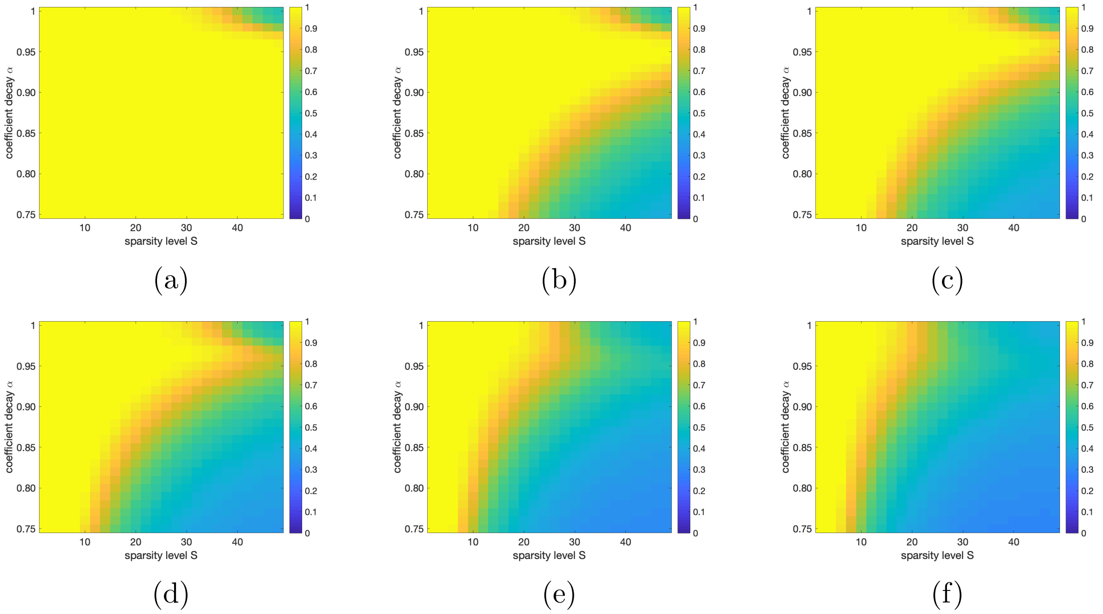
$$\psi_k = \left(1 - \frac{\varepsilon_k^2}{2}\right)\phi_k + \left(\varepsilon_k^2 - \frac{\varepsilon_k^4}{4}\right)^{\frac{1}{2}} z_k,$$

where $\varepsilon_k = \|\phi_k - \psi_k\|_2$ and z_k some unit random vector such that $\langle \phi_k, z_k \rangle = 0$.

In the following we conduct various experiments, in order to show how the degree of added perturbation (varying ε), the decay of the signal coefficients, noise as well as the properties of the signal generating dictionary influence the percentage of correctly recovered atoms. Further, we also run some experiments, comparing the performance of OMP with that of thresholding. For all our experiments we use $N = 1000$ signals and for the perturbed dictionary Ψ we consider the case where all atoms ψ_k are equally perturbed, meaning, we have $\|\phi_k - \psi_k\|_2 = \varepsilon$ for all k.

Noiseless signals with Dirac-DCT and Dirac-DCT random dictionary:

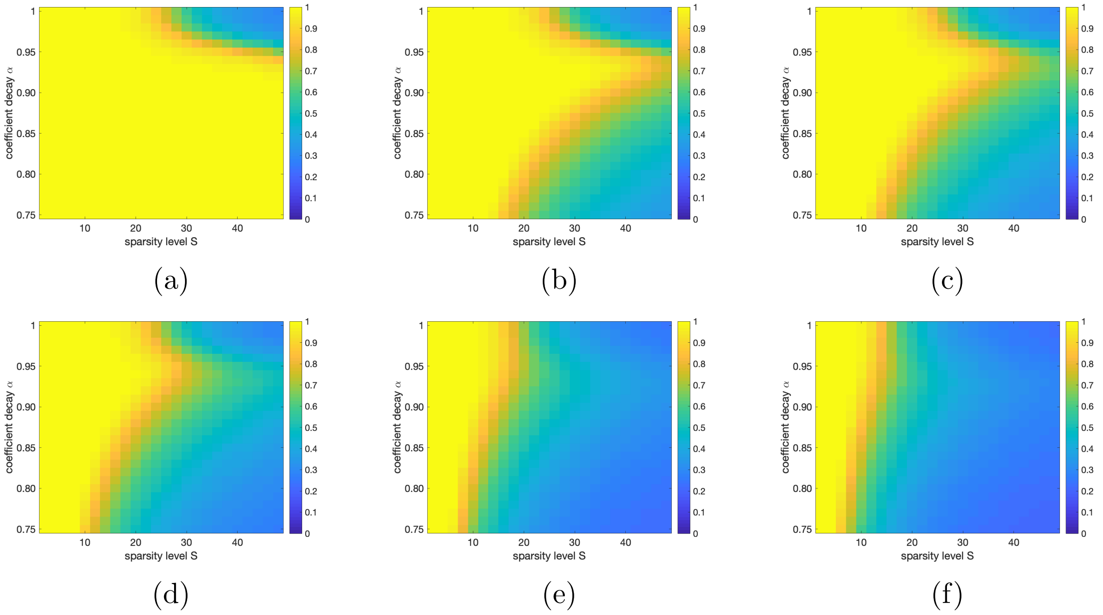
In our first experiment we consider noiseless signals as described above in the Dirac-DCT dictionary (Φ_{DD}) as well as the Dirac-DCT dictionary with additional $2d$ random vectors (Φ_{DDr}) and show how an increase of ε affects the recovery rate of OMP. For the coherence μ and the operator norm we have for the Dirac-DCT dictionary $\mu(\Phi_{DD}) = 0.125$ and $\|\Phi_{DD}\|_{2,2} = 1.42$, and for the Dirac-DCT random dictionary $\mu(\Phi_{DDr}) = 0.366$ and $\|\Phi_{DDr}\|_{2,2} = 2.74$. For the perturbation dictionaries Z_{DD} and Z_{DDr}, we have $\|Z_{DD}\|_{2,2} = 2.34$ and $\|Z_{DDr}\|_{2,2} = 2.95$.
Figure 6.1 shows the percentage of correctly recovered atoms via OMP with perturbations $\Psi_{DD} = \Phi_{DD}A + Z_{DD}W$ of the signal generating dictionary Φ_{DD} for various sparsity levels and decay parameters of the signal coefficients. From the results we can see that even very small distances between the generating and the perturbed dictionary ($\varepsilon = 0.05$) can cause a strong decrease in the recoverability of OMP. Further, while signal coefficients exhibiting more decay ensure the recovery of all correct atoms in case where we have the signal generating dictionary ($\varepsilon = 0$), this holds no longer true in case we only have a perturbation of it. Especially in the case of larger perturbations ($\varepsilon = 0.5$) and strongly decaying coefficients, OMP is only able to recover the full support for very sparse (S small) signals.
Figure 6.2 shows the equivalent results for OMP with perturbed dictionaries $\Psi_{DDr} = \Phi_{DDr}A + Z_{DDr}W$ and noiseless signals generated using Φ_{DDr}. Comparing the results in Figure 6.1 and Figure 6.2 we can clearly observe the better performance of OMP for the very well-behaved Dirac-DCT dictionary Φ_{DD}.



(a) (b) (c)

(d) (e) (f)

Figure 6.1: Percentage of correctly recovered atoms via OMP with perturbed dictionaries Ψ_{DD} with $\varepsilon = 0$ (a), $\varepsilon = 0.05$ (b), $\varepsilon = 0.1$ (c), $\varepsilon = 0.2$ (d), $\varepsilon = 0.4$ (e) and $\varepsilon = 0.5$ (f), for noiseless signals with generating dictionary Φ_{DD} and various sparsity levels and coefficient decay parameters.



(a) (b) (c)

(d) (e) (f)

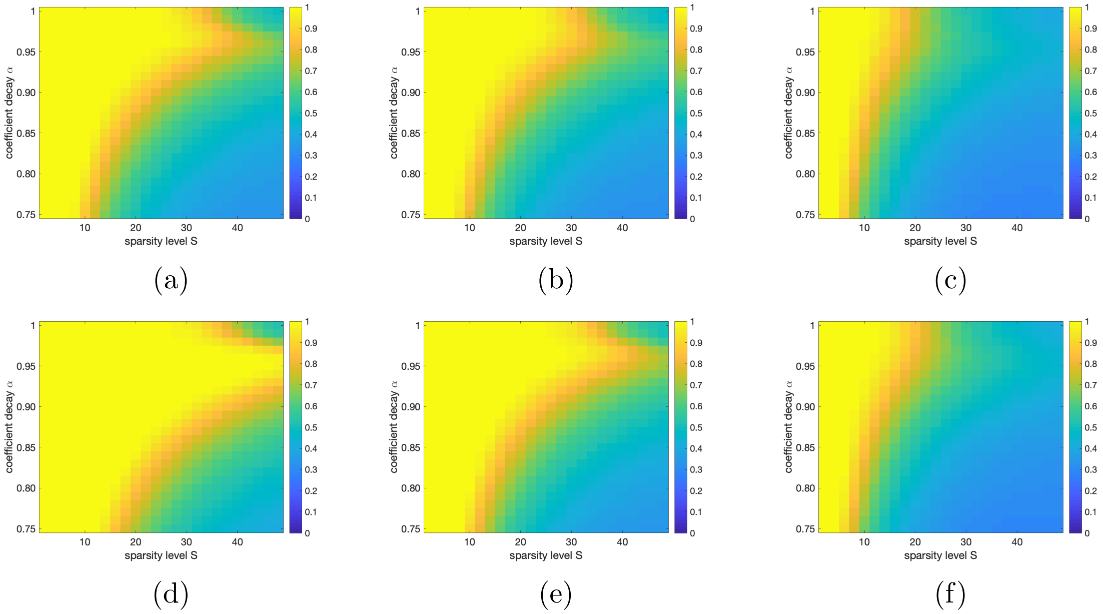
Figure 6.2: Percentage of correctly recovered atoms via OMP with perturbed dictionaries Ψ_{DDr} with $\varepsilon = 0$ (a), $\varepsilon = 0.05$ (b), $\varepsilon = 0.1$ (c), $\varepsilon = 0.2$ (d), $\varepsilon = 0.4$ (e) and $\varepsilon = 0.5$ (f), for noiseless signals with generating dictionary Φ_{DDr} and various sparsity levels and coefficient decay parameters.

Noisy signals with Dirac-DCT and Dirac-DCT random dictionary:

In our second experiment we consider noisy signals in the Dirac-DCT dictionary Φ_{DD} as well as the Dirac-DCT random dictionary Φ_{DDr}. For that, we additionally draw N noise vectors to create the signals. The noise variances are chosen to be $\rho^2 = \frac{1}{256d}$ and $\rho^2 = \frac{1}{16d}$, which correspond to signal to noise ratios (SNR) of 256 and 16, respectively. Figure 6.3 shows the obtained results for perturbed dictionaries $\Psi_{DD} = \Phi_{DD}A + Z_{DD}W$ of the signal generating dictionary Φ_{DD}, for noisy signals with SNR=16 in (a,b,c) and SNR=256 in (d,e,f), and for distances $\varepsilon = 0$ (a,d), $\varepsilon = 0.2$ (b,e) and $\varepsilon = 0.5$ (c,f). Comparing these results with the noiseless case in Figure 6.1, we can clearly observe the effect of noise in the perturbation free case ($\varepsilon = 0$). This decrease in the number of correctly recovered atoms occurs as in case of noisy signals we are only able to recover atoms corresponding to signal coefficients which are above the noise level. In case where we have to deal with perturbations of the generating dictionary ($\varepsilon > 0$), these effects are combined with the ones of the perturbation, resulting in an additional restriction of the range of parameters for which OMP performs well.

Figure 6.4 shows the equivalent results for perturbed dictionaries $\Psi_{DDr} = \Phi_{DDr}A + Z_{DDr}W$ of the signal generating dictionary Φ_{DDr}. Comparing the results in Figure 6.3 and Figure 6.4, we can again see the better performance of OMP for the very well-behaved dictionary Φ_{DD}.

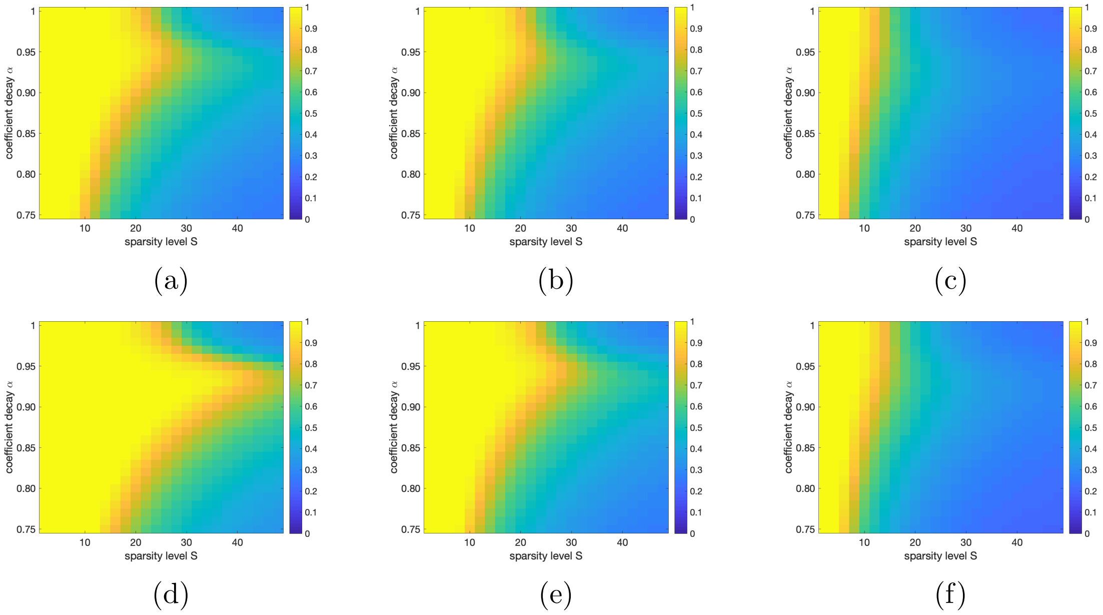
Another interesting observation is that the percentage of correctly recovered atoms in case of noiseless signals and $\varepsilon = 0.2$ is almost the same as for noisy signals with SNR=16 in the perturbation-free case, $\varepsilon = 0$. Similarly, for the noisy case with SNR=256 and $\varepsilon = 0$ we have almost the same results as for the noiseless case with $\varepsilon = 0.05$. For conciseness, we summarised the respective results in Figure 6.5.



(a) (b) (c)

(d) (e) (f)

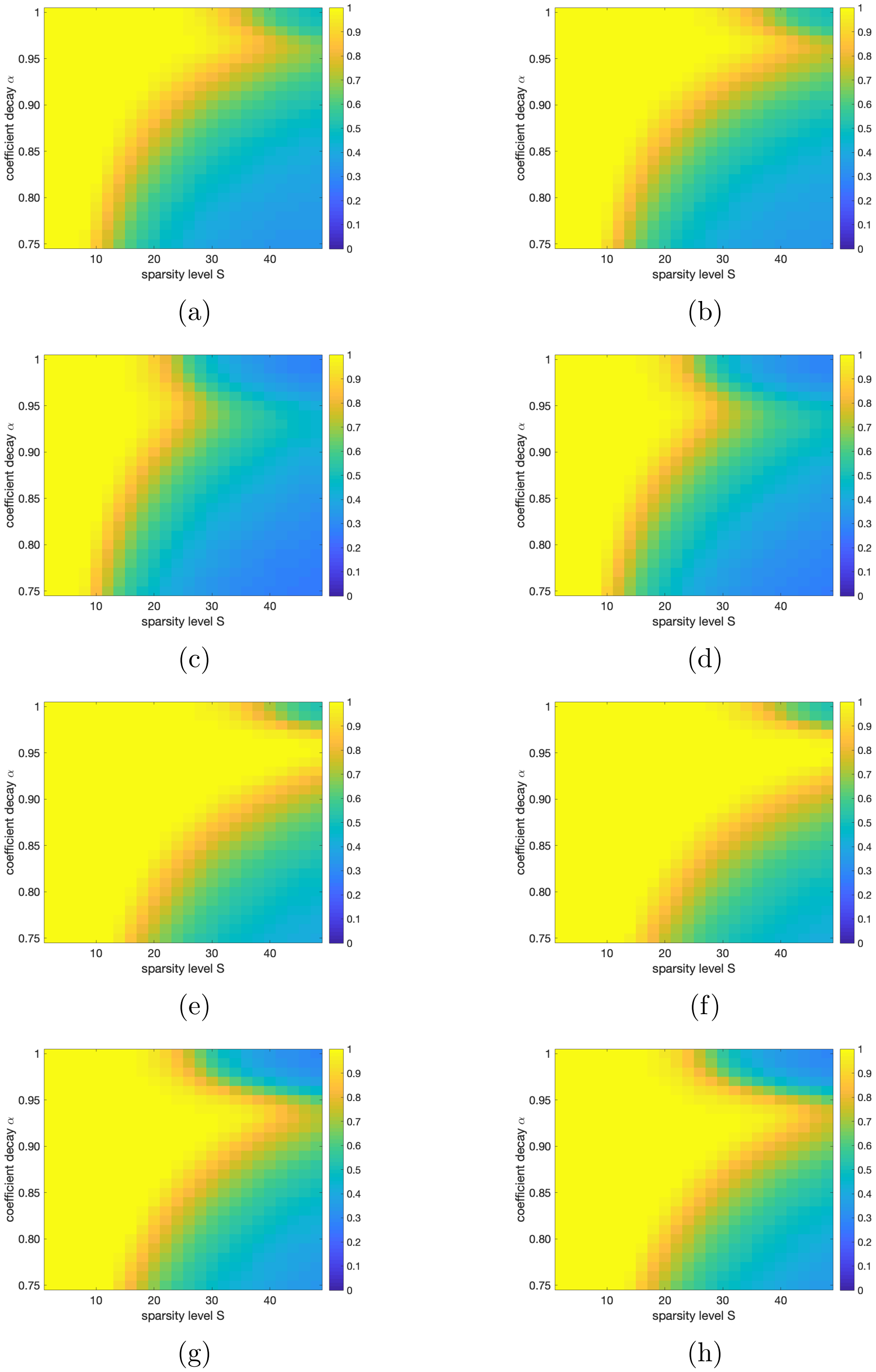
Figure 6.3: Percentage of correctly recovered atoms via OMP with perturbed dictionaries Ψ_{DD} with $\varepsilon = 0$ (a,d), $\varepsilon = 0.2$ (b,e) and $\varepsilon = 0.5$ (c,f), for signals in Φ_{DD} which are contaminated with Gaussian noise corresponding to SNR=16 (a,b,c) and SNR=256 (d,e,f), for various sparsity levels and coefficient decay parameters.



(a) (b) (c)

(d) (e) (f)

Figure 6.4: Percentage of correctly recovered atoms via OMP with perturbed dictionaries Ψ_{DDr} with $\varepsilon = 0$ (a,d), $\varepsilon = 0.2$ (b,e) and $\varepsilon = 0.5$ (c,f), for signals in Φ_{DDr} which are contaminated with Gaussian noise corresponding to SNR=16 (a,b,c) and SNR=256 (d,e,f), for various sparsity levels and coefficient decay parameters.



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

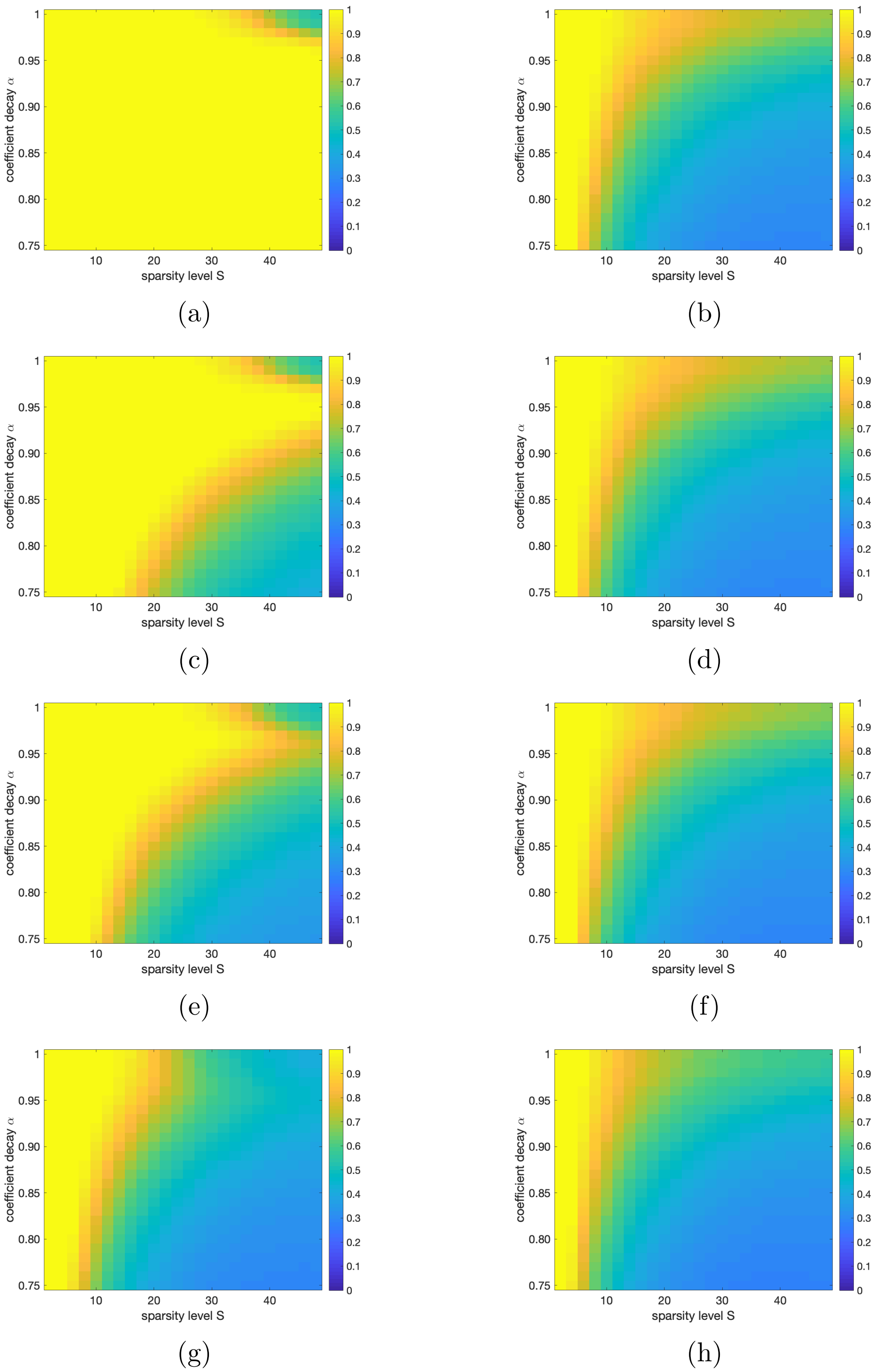
Figure 6.5: Connection between SNR and perturbations. Percentage of correctly recovered atoms via OMP for noisy signals with SNR=16 in Ψ_{DD} (a) and Ψ_{DDr} (c); for noisy signals with SNR=256 in Ψ_{DD} (e) and Ψ_{DDr} (g); for noiseless signals in Ψ_{DD} (b,f) and Ψ_{DDr} (d,h) with $\varepsilon = 0.2$ (b,d) and $\varepsilon = 0.05$ (f,h).

Comparison OMP and thresholding:

In our last experiment we compare the success rates of OMP with those of thresholding for noiseless signals in the Dirac-DCT dictionary Φ_{DD}. Similar to previous experiments we compare the percentage of correctly recovered atoms for various distances ε between the generating dictionary Φ_{DD} and a perturbation of it $\Psi_{DD} = \Phi_{DD}A + Z_{DD}W$. Further, we also give a comparison of how often OMP and thresholding are able to recover the full support of the corresponding signals.

The results in Figure 6.6 show the percentage of correctly recovered atoms via OMP (a,c,e,g) and thresholding (b,d,f,h) with Ψ_{DD} and signals using Φ_{DD}. The considered distances between the generating dictionary Φ_{DD} and Ψ_{DD} are $\varepsilon = 0$ (a,b), $\varepsilon = 0.05$ (c,d), $\varepsilon = 0.2$ (e,f) and $\varepsilon = 0.5$ (g,h). From the results we can clearly see the much better performance of OMP, especially for smaller ε. However, apart from the fact that the range of parameters for which thresholding performs well is very limited, thresholding seems to be more stable. A particularly interesting case is the one in (g) and (h). While we have a very huge gap between the performances of OMP and thresholding for $\varepsilon = 0$, comparing the percentage of correctly recover atoms for $\varepsilon = 0.5$, this gap closes more and more.

Finally, we also wanted to compare how often OMP and thresholding are able to recover the full support of our signals. In particular, Figure 6.7 shows the results which we obtained by counting for each pair (S, α) how often OMP (a,c,e,g) and thresholding (b,d,f,h) recovered the full support of the corresponding signals for distances $\varepsilon = 0$ (a,b), $\varepsilon = 0.05$ (c,d), $\varepsilon = 0.2$ (e,f) and $\varepsilon = 0.5$ (g,h). Again, we can see the much better performance of OMP which for increasing ε almost decreases to the one of thresholding.



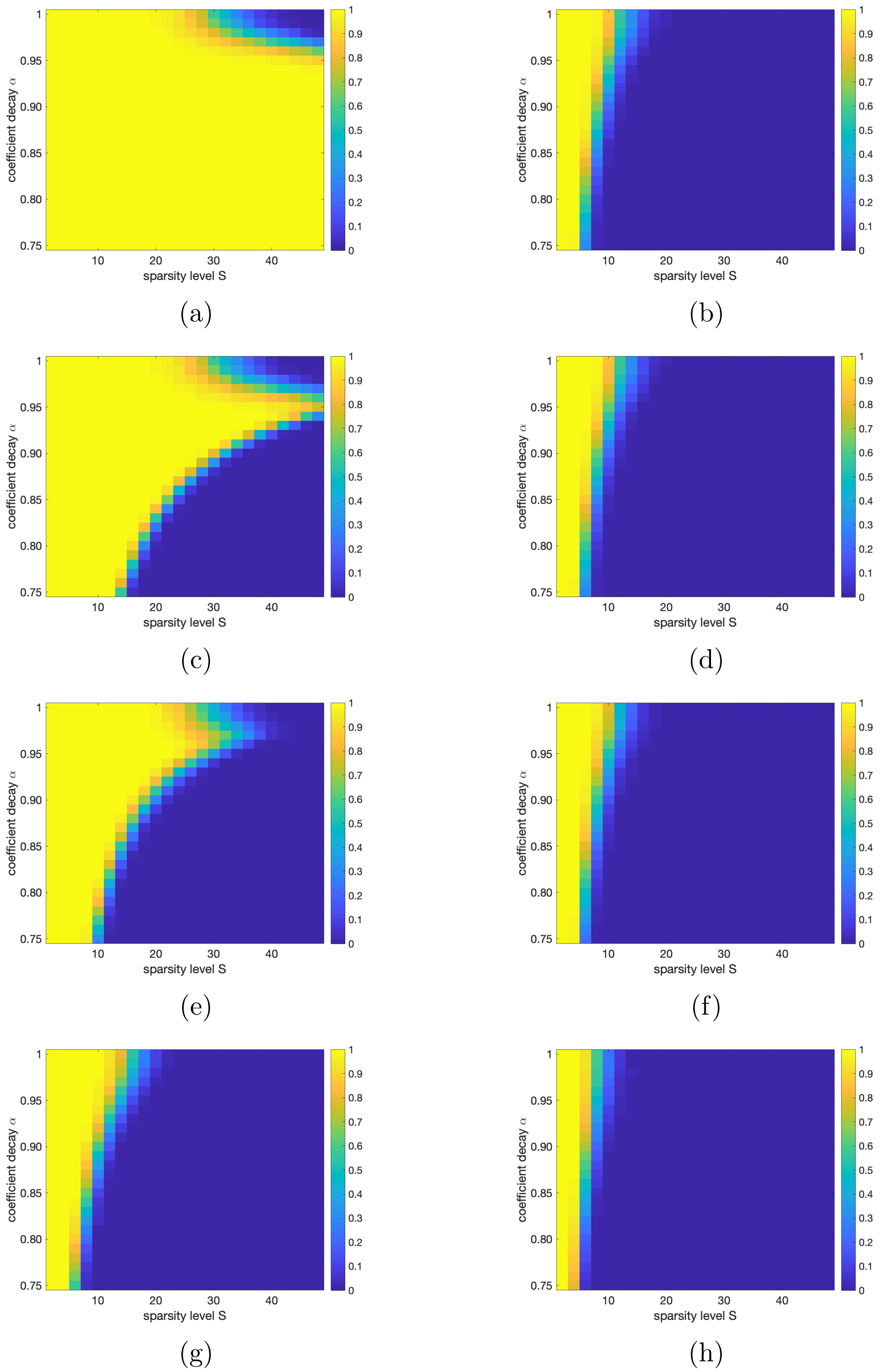
(a) (b)

(c) (d)

(e) (f)

(g) (h)

Figure 6.6: Percentage of correctly recovered atoms via OMP (a,c,e,g) and thresholding (b,d,f,h) with perturbed dictionaries Ψ_{DD} with $\varepsilon = 0$ (a,b), $\varepsilon = 0.05$ (c,d), $\varepsilon = 0.2$ (e,f) and $\varepsilon = 0.5$ (g,h), for noiseless signals with generating dictionary Φ_{DD} and various sparsity levels and coefficient decay parameters.



(a) (b)

(c) (d)

(e) (f)

(g) (h)

Figure 6.7: Percentage of correctly recovered supports via OMP (a,c,e,g) and thresholding (b,d,f,h) with perturbed dictionaries Ψ_{DD} with $\varepsilon = 0$ (a,b), $\varepsilon = 0.05$ (c,d), $\varepsilon = 0.2$ (e,f) and $\varepsilon = 0.5$ (g,h), for noiseless signals with generating dictionary Φ_{DD} and various sparsity levels and coefficient decay parameters.

6.6 Discussion

In this chapter we have presented average case results for OMP for the case where we do not have the signal generating dictionary but only a perturbed version of it. We have seen that even in the noiseless case, such perturbations can cause great difficulties as the conditions which ensure support recovery include a term which can grow very fast. In particular, signal properties ensuring (full-) support recovery in the case where we have the signal generating dictionary may provoke the opposite in the presence of perturbations. In case of noisy signals, an additional term has been added to the recovery conditions, causing further limitations of the range of parameters for which OMP performs well. We also gained interesting insights when comparing OMP with thresholding. From the theoretical results we have seen why thresholding performs that worse compared to OMP and that a similar term as the one which causes problems for thresholding also occurs within the recovery conditions of OMP if we have to deal with perturbations of the signal generating dictionary. From this we see, while OMP is a great sparse approximation algorithm in the perturbation-free case, its performance can be quite limited in the presence of perturbations. These observations could for example be used to accelerate dictionary learning algorithms which use OMP for updating the sparse support. In particular, one could replace OMP by simple thresholding within the first iterations and keep it for later iterations. Hence, accelerating the dictionary learning in the early stages using thresholding while keeping a good precision by using OMP in the later stages.

Chapter 7

Conclusion & Outlook

In this thesis we have taken a closer look at different areas of dictionary learning and sparse representation modelling. After introducing the main concepts in Chapter 2, we studied the contractive behaviour of the Iterative Thresholding and K residual Means (ITKrM) algorithm in Chapter 3. In particular, we showed that one iteration of ITKrM is a contraction under much more relaxed conditions compared to existing results. In Chapter 4, we analysed situations where ITKrM does not recover the generating dictionary. This showed us that there seem to exist stable fixed points which are not equivalent to the generating dictionary and can be characterised as very coherent. Based on a closer inspection of these spurious fixed points, we developed a replacement strategy and a strategy to find good replacement candidates. With the help of these replacement candidates we further addressed the question how to automatically choose the sparsity level S and the dictionary size K. In Chapter 5 we investigated the application of the adaptive version of ITKrM (aITKrM) together with an adaptive version of OMP (aOMP) to reconstruct MR images from highly undersampled data. By conducting various experiments we saw that the choice of the sparsity level S as well as the dictionary size K is non-trivial and strongly data dependent. However, using aITKrM and aOMP, S and K were no longer needed as input-parameter but optimally determined during the iterative reconstruction. Finally, Chapter 6 was devoted to the question how sparse approximation algorithms perform in case the given input dictionary is not the signal generating dictionary itself but a perturbed version of it. For that, we provided average case results for OMP in presence of perturbations of the generating dictionary and compared its performance with the one of simple thresholding.

While we gained a lot of interesting insights, unfortunately some questions also remained unanswered. For instance, in the convergence results of ITKrM we wanted to replace a very limiting condition on the signal coefficients by a more general one. In particular, we have that the convergence radius of ITKrM decreases with the dynamic

range of the coefficients. For that, in order to overcome large dynamic ranges, while at the same time removing the requirement of knowing the exact sparsity level S, we tried to extend the results to the case where we only assume a gap between the coefficients c_S and $c_{S\pm T}$, for some $T > 0$. Unfortunately we did not succeed as we ended up with norm terms which we did not get small enough.
Within the contraction conditions of ITKrM we also tried to reduce all $\log K$ to $\log S$ factors. While we were able to remove some of the $\log K$ factors by using the idea that an updated atom is only affected by the error originating from the failure of thresholding if its corresponding index is within the original support or if it is not within the original support but in the thresholded support, some of the $\log K$ factors remained due to a union bound over all indices which we could not replace.
Another interesting question that is still open concerns the theoretical analysis of another adaptive sparse approximation algorithm - Adaptive Pursuit - that has been shown to perform very well in numerical experiments. This algorithm works similar to aOMP however, it is able to add and remove several atoms at a time. More precisely, in each iteration it adds all atoms for which the residual inner product is larger than some predefined threshold times the norm of the residual and removes those for which the corresponding signal coefficient is below this critical value. While we were able to provide upper and lower bounds for thresholds ensuring that in each iteration at least one correct atom is added, we failed when proving that correct atoms are kept and erroneously picked ones are removed. The reason for this was that we not only had to upper and lower bound the size of the coefficients but also the norm term of the residual which occurs within the threshold. This finally led to bounds which were too restrictive.

Other interesting research topics would be the following. While the contraction theorem in Chapter 3 is a large improvement over existing results, it is however only valid for one iteration. Therefore, one interesting problem to be addressed is to prove that the updated dictionary inherits from the current dictionary estimate the properties that are required for being a contraction hence, ensuring convergence of ITKrM on a much larger area.
The results in Chapter 5 have clarified the importance of the adaptive choice of the sparsity level and the number of dictionary atoms. While aITKrM and aOMP have been shown to be significantly faster compared to the well-established K-SVD and OMP algorithms, they could be further accelerated and optimised. In particular, the underlying nature of ITKrM offers the possibility to transfer the calculations on a GPU and exploit parallelisation as it can process the patches sequentially. Further improvements in terms of computational time could also be expected from a more computationally efficient implementation of aITKrM and aOMP, e.g. by extending aOMP to use the Cholesky decomposition as in [54]. With such modifications, they could become an even bigger competitor for deep learning-based methods for medical imaging applications.

Another interesting question would be the average case performance of other sparse approximation algorithms in the presence of perturbations of the generating dictionary. Such results would be particularly interesting for developing or improving dictionary learning algorithms.

However, these are only a few of many other interesting and open problems to be answered in this very exciting field of research.

Appendix

A.1 Proof of Proposition 3.4

Here we prove Proposition 3.4, which was used within the proof of Lemma 3.5 in Chapter 3 to deal with sums of dependent random variables. For this, we need the following simplified version of Freedman's inequality.

Theorem A.1. (Freedman, [23]). *Let $X_0, \dots, X_S$ be a martingale sequence with bounded differences, that is $|X_k - X_{k-1}| \leq c$ almost surely for each k. Moreover, let the predictable quadratic variation $\langle X \rangle_S = \sum_{k=1}^{S} \mathbb{E}\left[(X_k - X_{k-1})^2 \middle| \mathcal{F}_{k-1}\right]$ be bounded by b. Then for all $t > 0$*

$$\mathbb{P}\left(X_S - X_0 \geq t\right) \leq \exp\left(-\frac{t^2}{2(ct+b)}\right).$$

Proposition 3.4. *Let $v \in \mathbb{R}^K$ be a vector, $I = (i_1, \dots, i_S)$ be a sequence of length S obtained by sampling from $\mathbb{K} = \{1, \dots, K\}$ without replacement, ε with values in $\{-1,1\}^S$ a Rademacher vector independent from I and $c \in \mathbb{R}^S$ a scaling vector. Then for any $t \geq 0$,*

$$\mathbb{P}\left(\big|\sum_{k=1}^{S} c_k \varepsilon_k v_{i_k}\big| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\big(\|c\|_\infty \|v\|_\infty t + \|c\|_2^2 \|v\|_2^2/(K-S)\big)}\right).$$

Proof. We will use Theorem A.1 on an appropriately constructed martingale. Let $I = (i_1, \dots, i_S)$, be the random vector obtained by sampling from $\mathbb{K} = \{1, \dots, K\}$ without replacement, that is, I is drawn uniformly at random from the set

$$\Omega := \{\omega \in \mathbb{K}^S : \omega_i \neq \omega_j \text{ for } i \neq j\}.$$

We equip Ω with the σ-algebra $\mathcal{F} := \mathcal{P}(\Omega)$ and the point measure $\mathbb{P}(\{\omega\}) := |\Omega|^{-1}$, to get the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We also set $\Delta := \{-1,1\}^S$, equip it with the σ-algebra $\mathcal{A} := \mathcal{P}(\Delta)$, the point measure $\mathbb{Q}(\{\delta\}) = 2^{-S}$ and define the product space $(\Omega \times \Delta, \mathcal{F} \otimes \mathcal{A}, \mathbb{P} \otimes \mathbb{Q})$. On Ω we define the filtration $\{\emptyset, \Omega\} = \mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}_S = \mathcal{F}$, where $\mathcal{F}_k$ is the σ-algebra induced by $\omega_1, \dots, \omega_k$. To be exact, we define the random variables

$$i_j : \Omega \longrightarrow \mathbb{K} \subseteq \mathbb{R} \quad \text{with} \quad i_j(\omega) := \omega_j,$$

and set $\mathcal{F}_k := \sigma(i_1, \dots, i_k)$ for all $1 \leq k \leq S$.

Since we also want to condition on the signs $\delta \in \Delta$, we define the random variables

$$\varepsilon_j : \Delta \longrightarrow \{-1,1\} \subseteq \mathbb{R} \quad \text{with} \quad \varepsilon_j(\delta) := \delta_j,$$

and the corresponding filtration $\{\emptyset, \Delta\} = \mathcal{A}_0 \subseteq \mathcal{A}_1 \subseteq \cdots \subseteq \mathcal{A}_S = \mathcal{A}$, by setting $\mathcal{A}_k = \sigma(\varepsilon_1, \ldots, \varepsilon_k)$. On the product space $\Omega \times \Delta$ we then get the filtration $\mathcal{F}_k \otimes \mathcal{A}_k$. Next we define the bounded random variables

$$X_k : \Omega \times \Delta \longrightarrow \mathbb{R}, \quad \text{with} \quad X_k(\omega, \delta) = \sum_{j=1}^{k} c_j \varepsilon_j(\delta) v_{i_j(\omega)}.$$

The random variables X_k form a martingale sequence with resp. to the filtration $\mathcal{F}_k \otimes \mathcal{A}_k$, since by independence of ε_k to $\mathcal{F} \otimes \mathcal{A}_{k-1}$ we have

$$\begin{aligned}
\mathbb{E}[X_k - X_{k-1} | \mathcal{F}_{k-1} \otimes \mathcal{A}_{k-1}] &= \mathbb{E}\big[\mathbb{E}[X_k - X_{k-1} | \mathcal{F} \otimes \mathcal{A}_{k-1}]\big|\mathcal{F}_{k-1} \otimes \mathcal{A}_{k-1}\big] \\
&= \mathbb{E}\big[\mathbb{E}[c_k \varepsilon_k v_{i_k} | \mathcal{F} \otimes \mathcal{A}_{k-1}]\big|\mathcal{F}_{k-1} \otimes \mathcal{A}_{k-1}\big] \\
&= c_k \mathbb{E}\left[v_{i_k} \mathbb{E}[\varepsilon_k | \mathcal{F} \otimes \mathcal{A}_{k-1}] | \mathcal{F}_{k-1} \otimes \mathcal{A}_{k-1}\right] \\
&= c_k \mathbb{E}\left[v_{i_k} \mathbb{E}[\varepsilon_k] | \mathcal{F}_{k-1} \otimes \mathcal{A}_{k-1}\right] = 0.
\end{aligned}$$

The sum we want to estimate is X_S with expectation $\mathbb{E}(X_S) = 0 = X_0$. Further we have $|X_k - X_{k-1}| = |c_k \varepsilon_k v_{i_k}| \leq \|c\|_\infty \|v\|_\infty$ as well as $(X_k - X_{k-1})^2 = c_k^2 v_{i_k}^2$, so we can bound the predictable quadratic variation as

$$\begin{aligned}
\langle X \rangle_S &= \sum_{k=1}^{S} \mathbb{E}\left[(X_k - X_{k-1})^2 \big| \mathcal{F}_{k-1} \otimes \mathcal{A}_{k-1}\right] \\
&= \sum_{k=1}^{S} \mathbb{E}\left[c_k^2 v_{i_k}^2 \big| \mathcal{F}_{k-1} \otimes \mathcal{A}_{k-1}\right] = \sum_{k=1}^{S} c_k^2 \sum_{\ell \notin \{i_1, \ldots, i_{k-1}\}} v_\ell^2 \frac{1}{K - k + 1} \leq \|c\|_2^2 \frac{\|v\|_2^2}{K - S}.
\end{aligned}$$

The final result follows using the symmetry of X_S. □

A.2 Proof of Proposition 6.2

Here we prove Proposition 6.2, which concerns the norm of a random subvector and was used to prove Theorem 6.1 in Chapter 6. For this, we again need Theorem A.1.

Proposition 6.2. *Let $v \in \mathbb{R}^K$ be a vector, I a subset chosen uniformly at random among all subsets of size S, and v_I the restriction of v to the subset I, then for any $t \geq 0$,*

$$\mathbb{P}\left(\|v_I\|_2^2 \geq \tfrac{S}{K}\|v\|_2^2 + t\right) \leq \exp\left(-\frac{t^2}{2\left(\|v\|_\infty^2 t + \frac{S}{K}\|v\|_4^4\right)}\right).$$

Proof. Similar to Proposition 3.4 we let $I = (i_1, \ldots, i_S)$ be the random vector obtained by sampling from $\mathbb{K} = \{1, \ldots, K\}$ without replacement and define $(\Omega, \mathcal{F}, \mathbb{P})$, i_j and $\mathcal{F}_k$

as above. For notational convenience, later on, we also define the set-valued functions $\mathcal{I}_k^c$ where $\mathcal{I}_k^c(\omega) := \mathbb{K}/\{i_1(\omega), \dots, i_k(\omega)\}$.

Next we define the bounded random variable

$$X : \Omega \longrightarrow \mathbb{R}, \quad \text{with} \quad X(\omega) = \sum_{j=1}^{S} v_{i_j(\omega)}^2$$

and set $X_k := \mathbb{E}[X|\mathcal{F}_k]$. By construction $\mathbb{E}[X_{k+1}|\mathcal{F}_k] = X_k$ and so $X_0, X_1, \dots, X_S$ is a martingale with $X_0 = \mathbb{E}(X)$ and $X_S = X$. To apply Theorem A.1 we first have to bound $Y_k := X_k - X_{k-1}$. We have

$$\begin{aligned} Y_k &= X_k - X_{k-1} \\ &= \sum_{j=1}^{S} \mathbb{E}\big[v_{i_j}^2 \big| \sigma(i_1, \dots, i_k)\big] - \sum_{j=1}^{S} \mathbb{E}\Big[v_{i_j}^2 \Big| \sigma(i_1, \dots, i_{k-1})\Big]. \end{aligned}$$

Since we have $\mathbb{E}\big[v_{i_j}^2 \big| \sigma(i_1, \dots, i_k)\big] = v_{i_j}^2$ for $j \le k$ the expression above reduces to

$$Y_k = \sum_{j=k}^{S} \mathbb{E}\big[v_{i_j}^2 \big| \sigma(i_1, \dots, i_k)\big] - \sum_{j=k}^{S} \mathbb{E}\Big[v_{i_j}^2 \Big| \sigma(i_1, \dots, i_{k-1})\Big].$$

For $j > k$ on the other hand, we have for $\ell \in \mathbb{K}$,

$$\mathbb{P}\left(i_j = \ell | \sigma(i_1, \dots, i_k)\right) = \begin{cases} \frac{1}{K-k}, & \ell \notin \{i_1, \dots, i_k\}, \\ 0, & \text{else}, \end{cases}$$

thus,

$$\mathbb{E}\big[v_{i_j}^2 \big| \sigma(i_1, \dots, i_k)\big] = \sum_{\ell \in \mathbb{K}} v_\ell^2 \, \mathbb{P}\left(i_j = \ell | \sigma(i_1, \dots, i_k)\right) = \sum_{\ell \notin \{i_1, \dots, i_k\}} \frac{v_\ell^2}{K-k} = \frac{\|v_{\mathcal{I}_k^c}\|_2^2}{K-k},$$

so we can further simplify

$$\begin{aligned} Y_k &= v_{i_k}^2 + \frac{S-k}{K-k}\left(\|v_{\mathcal{I}_k^c}\|_2^2\right) - \frac{S-k+1}{K-k+1}\left(\|v_{\mathcal{I}_{k-1}^c}\|_2^2\right) \\ &= v_{i_k}^2 + \frac{S-k}{K-k}\left(\|v_{\mathcal{I}_{k-1}^c}\|_2^2 - v_{i_k}^2\right) - \frac{S-k+1}{K-k+1}\left(\|v_{\mathcal{I}_{k-1}^c}\|_2^2\right) \\ &= v_{i_k}^2\left(1 - \frac{S-k}{K-k}\right) + \|v_{\mathcal{I}_{k-1}^c}\|_2^2\left(\frac{S-k}{K-k} - \frac{S-k+1}{K-k+1}\right) \\ &= v_{i_k}^2 \frac{K-S}{K-k} - \|v_{\mathcal{I}_{k-1}^c}\|_2^2 \frac{K-S}{(K-k)(K-k+1)} \\ &= \frac{K-S}{K-k}\left(v_{i_k}^2 - \frac{\|v_{\mathcal{I}_{k-1}^c}\|_2^2}{K-k+1}\right). \end{aligned} \tag{A.2.1}$$

Since $\|v_{\mathcal{I}^c_{k-1}}\|_2^2 \le (K-k+1)\|v\|_\infty^2$ we can also bound the modulus of the difference in the bracket above by $\|v\|_\infty^2$ and further get that $|Y_k| = |X_k - X_{k-1}| \le \|v\|_\infty^2$.

Next we bound the predictable quadratic variation $\langle X\rangle_S$. Using the convenient expression for Y_k from (A.2.1), we have

$$\begin{aligned}\langle X\rangle_S &= \sum_{k=1}^{S} \mathbb{E}\left[Y_k^2 \middle| \mathcal{F}_{k-1}\right] \\ &= \sum_{k=1}^{S} \left(\frac{K-S}{K-k}\right)^2 \mathbb{E}\left[\left(v_{i_k}^2 - \frac{\|v_{\mathcal{I}^c_{k-1}}\|_2^2}{K-k+1}\right)^2 \middle| \mathcal{F}_{k-1}\right] \\ &= \sum_{k=1}^{S} \left(\frac{K-S}{K-k}\right)^2 \mathbb{E}\left[v_{i_k}^4 - 2v_{i_k}^2 \frac{\|v_{\mathcal{I}^c_{k-1}}\|_2^2}{K-k+1} + \frac{\|v_{\mathcal{I}^c_{k-1}}\|_2^4}{(K-k+1)^2} \middle| \mathcal{F}_{k-1}\right].\end{aligned}$$

Since $\mathcal{I}^c_{k-1} = \mathbb{K}/\{i_1,\dots,i_{k-1}\}$, the function $\|v_{\mathcal{I}^c_{k-1}}\|_2^2$ is $\mathcal{F}_{k-1}$-measurable and we get

$$\langle X\rangle_S = \sum_{k=1}^{S} \left(\frac{K-S}{K-k}\right)^2 \left(\mathbb{E}\left[v_{i_k}^4 \middle| \mathcal{F}_{k-1}\right] - 2\mathbb{E}\left[v_{i_k}^2 \middle| \mathcal{F}_{k-1}\right] \frac{\|v_{\mathcal{I}^c_{k-1}}\|_2^2}{K-k+1} + \frac{\|v_{\mathcal{I}^c_{k-1}}\|_2^4}{(K-k+1)^2}\right).$$

For $p = 2, 4$ we have

$$\mathbb{E}\left[v_{i_k}^p \middle| \mathcal{F}_{k-1}\right] = \sum_{\ell \notin \{i_1,\dots,i_{k-1}\}} v_\ell^p \frac{1}{K-k+1} = \frac{\|v_{\mathcal{I}^c_{k-1}}\|_p^p}{K-k+1},$$

which leads to

$$\begin{aligned}\langle X\rangle_S &= \sum_{k=1}^{S} \left(\frac{K-S}{K-k}\right)^2 \left(\frac{\|v_{\mathcal{I}^c_{k-1}}\|_4^4}{K-k+1} - \frac{\|v_{\mathcal{I}^c_{k-1}}\|_2^4}{(K-k+1)^2}\right) \\ &= \sum_{k=1}^{S} \frac{(K-S)^2}{(K-k)^2(K-k+1)^2} \left((K-k+1)\,\|v_{\mathcal{I}^c_{k-1}}\|_4^4 - \|v_{\mathcal{I}^c_{k-1}}\|_2^4\right).\end{aligned}$$

Since for any vector $z \in \mathbb{R}^d$ we have $\|z\|_4^4 \le \|z\|_2^4 \le d\|z\|_4^4$, we arrive at the final bound

$$\begin{aligned}\langle X\rangle_S &\le \sum_{k=1}^{S} \frac{(K-S)^2}{(K-k)(K-k+1)^2} \|v_{\mathcal{I}^c_{k-1}}\|_4^4 \le \sum_{k=1}^{S} \frac{(K-S)^2}{K-S+1} \cdot \frac{\|v\|_4^4}{(K-k)(K-k+1)} \\ &\le \|v\|_4^4 (K-S) \sum_{k=1}^{S} \frac{1}{K-k} - \frac{1}{K-k+1} \le \|v\|_4^4 \, \frac{S}{K}.\end{aligned}$$

□

A.3 Proof of Theorem 6.3

Here we state the proof of Theorem 6.3 which provides partial support recovery conditions for OMP in case of noisy signals.

Proof. We use the same strategy and abbreviations as in the proof of Theorem 6.1. In case of noisy signals the residuals are now of the form

$$\tilde{r}_{\bar{J}} = Q(\Psi_{\bar{J}})\tilde{y} = Q(\Psi_{\bar{J}})(\Phi_{\bar{I}} x_I + \eta) = r_{\bar{J}} + Q(\Psi_{\bar{J}})\eta.$$

A sufficient condition to ensure that OMP picks another correct atom within the next iteration is that we have for i, the index of the largest still missing coefficient,

$$\left|\langle \psi_{p(i)}, \tilde{r}_{\bar{J}} \rangle\right| > \max_{k \in R_i} \left|\langle \psi_{p(k)}, \tilde{r}_{\bar{J}} \rangle\right| . \tag{A.3.1}$$

Note that we have

$$\left|\langle \psi_{p(i)}, \tilde{r}_{\bar{J}} \rangle\right| \geq \left|\langle \psi_{p(i)}, r_{\bar{J}} \rangle\right| - \left|\langle \psi_{p(i)}, Q(\Psi_{\bar{J}})\eta \rangle\right| ,$$

and for all $k \in R_i$,

$$\left|\langle \psi_{p(k)}, \tilde{r}_{\bar{J}} \rangle\right| \leq \left|\langle \psi_{p(k)}, r_{\bar{J}} \rangle\right| + \left|\langle \psi_{p(k)}, Q(\Psi_{\bar{J}})\eta \rangle\right| ,$$

and hence, the condition in (A.3.1) is implied by ensuring that for i and all $k \in R_i$, we have

$$\left|\langle \psi_{p(i)}, r_{\bar{J}} \rangle\right| - \left|\langle \psi_{p(k)}, r_{\bar{J}} \rangle\right| \geq \left|\langle \psi_{p(i)}, Q(\Psi_{\bar{J}})\eta \rangle\right| + \left|\langle \psi_{p(k)}, Q(\Psi_{\bar{J}})\eta \rangle\right| . \tag{A.3.2}$$

Since the terms on the left hand side are the same as for the noiseless perfectly S-sparse case, we first bound the terms on the right hand side of (A.3.2). For that, we use the decomposition $\bar{J} = \bar{A}_i \cup \bar{G}$, where $\bar{A}_i := p(A_i)$ and $\bar{G} := p(G)$. Hence, we can write for any $k \notin J$,

$$\begin{aligned}
|\langle \psi_{p(k)}, Q(\Psi_{\bar{J}})\eta \rangle| &= |\langle \psi_{p(k)}, Q(\Psi_{\bar{J}}) \left[P(\Psi_{\bar{A}_i}) + Q(\Psi_{\bar{A}_i})\right] \eta \rangle| \\
&= |\langle \psi_{p(k)}, Q(\Psi_{\bar{A}_i})\eta \rangle - \langle \psi_{p(k)}, P(\Psi_{\bar{J}}) Q(\Psi_{\bar{A}_i})\eta \rangle| \\
&= |\langle \psi_{p(k)}, Q(\Psi_{\bar{A}_i})\eta \rangle - \langle \Psi_{\bar{J}}^{\star}\psi_{p(k)}, (\Psi_{\bar{J}}^{\star}\Psi_{\bar{J}})^{-1}\Psi_{\bar{J}}^{\star} Q(\Psi_{\bar{A}_i})\eta \rangle|.
\end{aligned} \tag{A.3.3}$$

Since $\left|\langle \psi_{p(j)}, Q(\Psi_{\bar{A}_i})\eta \rangle\right| = 0$ for all $j \in A_i$, we get for any $k \notin J$

$$\begin{aligned}
|\langle \psi_{p(k)}, Q(\Psi_{\bar{J}})\eta \rangle| &\leq |\langle \psi_{p(k)}, Q(\Psi_{\bar{A}_i})\eta \rangle| + \|\Psi_{\bar{J}}^{\star}\psi_{p(k)}\|_2 \cdot \|(\Psi_{\bar{J}}^{\star}\Psi_{\bar{J}})^{-1}\|_{2,2} \cdot \|\Psi_{\bar{G}}^{\star} Q(\Psi_{\bar{A}_i})\eta\|_2 \\
&\leq |\langle \psi_{p(k)}, Q(\Psi_{\bar{A}_i})\eta \rangle| + \|\Psi_{\bar{J}}^{\star}\psi_{p(k)}\|_2 \cdot \frac{\sqrt{|G|}}{1-\delta_{\bar{J}}} \cdot \max_{\ell \in G} |\langle \psi_{p(\ell)}, Q(\Psi_{\bar{A}_i})\eta \rangle| \\
&\leq \max_{k \notin A_i} |\langle \psi_{p(k)}, Q(\Psi_{\bar{A}_i})\eta \rangle| \left(1 + \frac{\sqrt{|G|}}{1-\delta_{\bar{J}}} \cdot \|\Psi_{\bar{J}}^{\star}\psi_{p(k)}\|_2\right) \\
&\leq \max_{k \notin A_i} |\langle Q(\Psi_{\bar{A}_i})\psi_{p(k)}, \eta \rangle| \left(1 + \frac{\sqrt{|G| \cdot |J|}}{1-\delta_{\bar{J}}} \cdot \bar{\mu}\right) .
\end{aligned} \tag{A.3.4}$$

In order to bound $|\langle Q(\Psi_{\bar{A}_i})\psi_{p(k)}, \eta\rangle|$ for all $k \notin A_i$ with high probability, we use the sub-Gaussian property of the noise vector η. In particular, for the marginals $\langle v_{i_k}, \eta\rangle$, with $v_{i_k} = Q(\Psi_{\bar{A}_i})\psi_{p(k)}$ and $\|v_{i_k}\|_2 \leq 1$, we have

$$\mathbb{P}\left(\left|\langle Q(\Psi_{\bar{A}_i})\psi_{p(k)}, \eta\rangle\right| \geq \theta_\eta\right) \leq 2\exp\left(-\frac{\theta_\eta^2}{2\rho^2}\right).$$

Using a union bound over all k and all possible subsets A_i, we obtain

$$\mathbb{P}\left(\exists A_i, k : \left|\langle Q(\Psi_{\bar{A}_i})\psi_{p(k)}, \eta\rangle\right| \geq \theta_\eta\right) \leq 2sK\exp\left(-\frac{\theta_\eta^2}{2\rho^2}\right).$$

Substituting this bound into (A.3.4) and setting $\theta_\eta = 2\rho\sqrt{n\log K}$, we have that except with probability $2sK^{1-2n}$,

$$|\langle\psi_{p(k)}, Q(\Psi_{\bar{J}})\eta\rangle| \leq 2\rho\sqrt{n\log K}\left(1 + \frac{\sqrt{|G|\cdot|J|}}{1-\delta_{\bar{J}}}\cdot\bar{\mu}\right) \qquad \text{for all} \quad k, \bar{J}.$$

Using that $|G| = |C| \cup |D| \leq 2t$ and $\delta_{\bar{J}} \leq \delta_{\bar{I}}(\Psi) \leq \frac{1}{2}$, we have that except with probability $2sK^{1-2n}$,

$$|\langle\psi_{p(k)}, Q(\Psi_{\bar{J}})\eta\rangle| \leq 2\rho\sqrt{n\log K}\left(1 + 2\sqrt{2t\cdot|J|}\cdot\bar{\mu}\right). \tag{A.3.5}$$

Next, we bound the terms on the left hand side of (A.3.2). For that, we use the bounds derived in Theorem 6.1 but with a slight tweak as we are interested in conditions ensuring only partial support recovery. In particular, as already mentioned in the proof of Theorem 6.1, in case $\bar{J}$ contains only part of the elements in $\bar{I}$, we might get better estimates by using only a crude bound for the norm term $\|\Psi^\star_{\bar{J}}\psi_{p(k)}\|_2$. For that, going back to (6.41) and replacing E by the bound $\sqrt{|J|}\bar{\mu}$, we have for $i \in G^c$, the index of the largest still missing coefficient, and any $k \in R_i$,

$$\begin{aligned}
&c_i^{-1}\gamma_{\min}^{-1}\left(\left|\langle\psi_{p(i)}, r_{\bar{J}}\rangle\right| - \left|\langle\psi_{p(k)}, r_{\bar{J}}\rangle\right|\right) \\
&\geq 1 - \frac{\gamma}{\gamma_{\min}}\beta \\
&\qquad - 4\frac{\gamma}{\gamma_{\min}}\cdot\varepsilon\sqrt{n\log K}\cdot\max\left\{2\nu_Z\frac{\|c_I\|_\infty}{c_i}\sqrt{n\log K}, \frac{\|c_I\|_2}{c_i}\sqrt{\frac{\|Z\|_{2,2}^2}{K-S}}\right\}(1+2|J|\bar{\mu}) \\
&\qquad\quad - 4\frac{\gamma}{\gamma_{\min}}\cdot\bar{\mu}\cdot\left(t + \sqrt{nt\log K}\left(\frac{\beta^2}{1-\beta^2}\right)^{\frac{1}{2}}\right)(1+2|J|\bar{\mu}),
\end{aligned} \tag{A.3.6}$$

except with probability $K(2K^{-2n} + 2K^{-n} + 216K^{-m})$.

Combining these estimates with the bound in (A.3.5), a sufficient condition ensuring the recovery of $p(i)$ before $p(k)$ with k in $\bar{R}_i$ in case of noisy signals is therefore

$$\begin{aligned}
&1 - \frac{\gamma}{\gamma_{\min}}\beta \\
&\quad > 4\frac{\gamma}{\gamma_{\min}} \cdot \varepsilon\sqrt{n\log K} \cdot \max\left\{2\nu_Z \frac{\|c_I\|_\infty}{c_i}\sqrt{n\log K}, \frac{\|c_I\|_2}{c_i}\sqrt{\frac{\|Z\|_{2,2}^2}{K-S}}\right\}(1+2|J|\bar{\mu}) \\
&\qquad + 2\frac{\gamma}{\gamma_{\min}} \cdot \bar{\mu} \cdot \left(2t + 2\sqrt{nt\log K}\left(\frac{\beta^2}{1-\beta^2}\right)^{\frac{1}{2}}\right)(1+2|J|\bar{\mu}) \\
&\qquad\quad + 4\frac{\rho}{\gamma_{\min} c_i}\sqrt{n\log K}\left(1 + 2\bar{\mu}\sqrt{2t|J|}\right).
\end{aligned}$$

Using that in the ℓ-th iteration $|J| \leq \ell$ and c_ℓ is the smallest possible largest missing coefficient, we have that OMP recovers only correct atoms within the first ℓ iterations, except with probability $K(4sK^{-2n} + 2K^{-n} + 216K^{-m})$, as long as

$$\begin{aligned}
&1 - \frac{\gamma}{\gamma_{\min}}\beta \\
&\quad \geq 4\frac{\gamma}{\gamma_{\min}} \cdot \varepsilon\sqrt{n\log K} \cdot \max\left\{2\nu_Z \frac{\|c_I\|_\infty}{c_\ell}\sqrt{n\log K}, \frac{\|c_I\|_2}{c_\ell}\sqrt{\frac{\|Z\|_{2,2}^2}{K-S}}\right\}(1+2\ell\bar{\mu}) \\
&\qquad + 4\frac{\gamma}{\gamma_{\min}} \cdot \bar{\mu} \cdot \left(t + \sqrt{nt\log K}\left(\frac{\beta^2}{1-\beta^2}\right)^{\frac{1}{2}}\right)(1+2\ell\bar{\mu}) \\
&\qquad\qquad + 4\frac{\rho}{\gamma_{\min} c_\ell}\sqrt{n\log K}\left(1 + 2\bar{\mu}\sqrt{2t\ell}\right).
\end{aligned}$$

□

A.4 Proof of Theorem 6.4

Here we state the proof of Theorem 6.4 which provides support recovery conditions for thresholding in case of noiseless perfectly S-sparse signals.

Proof. Throughout the proof we will use the short hands $\bar{\mu} := \max_{i\neq j}|\langle\psi_i,\psi_j\rangle|$ and $\nu_Z := \max_{i,j}|\langle\psi_i, z_j\rangle|$. Further, we assume w.l.o.g. that $\bar{I} = I = \{1,\dots,S\}$. Note that, for our result we only use Hoeffding's inequality where the expectation is only over the sign sequence σ and hence, independent of the permutation p.
In order to ensure that thresholding succeeds, this means, to ensure the recovery of all $i \in I$, we need to have

$$\min_{i\in I}|\langle\psi_i, y\rangle| > \max_{i\notin I}|\langle\psi_i, y\rangle|. \tag{A.4.1}$$

Using the decomposition

$$y = \Phi_I x_I = (\Psi\Gamma)_I x_I + (\Phi - \Psi\Gamma)_I x_I = (\Psi\Gamma)_I x_I + (Z\Lambda)_I x_I, \tag{A.4.2}$$

the inner product of a signal y with an atom ψ_i of the perturbed dictionary Ψ, can be expanded as

$$\begin{aligned} |\langle \psi_i, y\rangle| &= |\langle \psi_i, (\Psi\Gamma)_I x_I\rangle + \langle \psi_i, (Z\Lambda)_I x_I\rangle| \\ &= \big|\sum_{j\in I} \langle \psi_i, \psi_j\rangle \gamma_j x_j + \sum_{j\in I} \langle \psi_i, z_j\rangle \lambda_j x_j\big| \\ &= \big|\sum_{j\in I} \langle \psi_i, \psi_j\rangle \gamma_j \sigma_j c_j + \sum_{j\in I} \langle \psi_i, z_j\rangle \lambda_j \sigma_j c_j\big| \\ &= \big|c_i\gamma_i + \sigma_i \sum_{j\in I\setminus\{i\}} \langle \psi_i, \psi_j\rangle \gamma_j \sigma_j c_j + \sum_{j\in I} \langle \psi_i, z_j\rangle \lambda_j \sigma_j c_j\big|. \end{aligned} \tag{A.4.3}$$

Depending on the index i under consideration, we get the following bounds from below resp. above,

$$\begin{aligned} i\in I : |\langle \psi_i, y\rangle| &\geq c_i\gamma_i - \big|\sum_{j\in I\setminus\{i\}} \langle \psi_i, \psi_j\rangle \gamma_j \sigma_j c_j\big| - \big|\sum_{j\in I} \langle \psi_i, z_j\rangle \lambda_j \sigma_j c_j\big|, \\ i\notin I : |\langle \psi_i, y\rangle| &\leq \big|\sum_{j\in I} \langle \psi_i, \psi_j\rangle \gamma_j \sigma_j c_j\big| + \big|\sum_{j\in I} \langle \psi_i, z_j\rangle \lambda_j \sigma_j c_j\big|. \end{aligned}$$

Therefore, a sufficient condition for the recovery of I is that for all $i \in I$, we have

$$c_i\gamma_i > 2\cdot\max_k \big|\sum_{j\in I\setminus\{k\}} \langle \psi_k, \psi_j\rangle \gamma_j \sigma_j c_j\big| + 2\cdot\max_k \big|\sum_{j\in I} \langle \psi_k, z_j\rangle \lambda_j \sigma_j c_j\big|. \tag{A.4.4}$$

In order to bound the right hand side, we use Hoeffding's inequality. Hence, for the inner product with the perturbed dictionary Ψ, we obtain

$$\begin{aligned} \mathbb{P}\big(\big|\sum_{j\in I\setminus\{k\}} \gamma_j\sigma_j c_j \langle \psi_k, \psi_j\rangle\big| \geq t_1\big) &\leq 2\exp\left(-\frac{t_1^2}{2\sum_{j\in I\setminus\{k\}} |\langle \psi_k, \psi_j\rangle|^2 \gamma_j^2 c_j^2}\right) \\ &\leq 2\exp\left(-\frac{t_1^2}{2\bar{\mu}^2\|\Gamma_I c_I\|_2^2}\right), \end{aligned}$$

and similarly for the inner product with the perturbation dictionary Z,

$$\begin{aligned} \mathbb{P}\big(\big|\sum_{j\in I} \lambda_j\sigma_j c_j \langle \psi_k, z_j\rangle\big| \geq t_2\big) &\leq 2\exp\left(-\frac{t_1^2}{2\sum_{j\in I} |\langle \psi_k, z_j\rangle|^2 \lambda_j^2 c_j^2}\right) \\ &\leq 2\exp\left(-\frac{t_1^2}{2\nu_Z^2\|\Lambda_I c_I\|_2^2}\right). \end{aligned}$$

Using a union bound over all indices k, we have that except with probability $2K^{1-2n_1}$

$$\max_k \big| \sum_{j \in I \setminus \{k\}} \gamma_j \sigma_j c_j \langle \psi_k, \psi_j \rangle \big| \leq 2\bar{\mu} \cdot \|\Gamma_I c_I\|_2 \sqrt{n_1 \log K}, \tag{A.4.5}$$

and except with probability $2K^{1-2n_2}$

$$\max_k \big| \sum_{j \in I} \lambda_j \sigma_j c_j \langle \psi_k, z_j \rangle \big| \leq 2\nu_Z \cdot \|\Lambda_I c_I\|_2 \sqrt{n_2 \log K}. \tag{A.4.6}$$

Defining $\gamma_{\min} = \min_k \gamma_k$, $\varepsilon_{\min} = \min_k \varepsilon_k$, $\gamma = \max_k \gamma_k$ and $\varepsilon = \max_k \varepsilon_k$, we have for $\varepsilon \leq 1$,

$$\begin{aligned} \gamma_i &= \frac{2}{2 - \varepsilon_i^2} \geq \frac{2}{2 - \varepsilon_{\min}^2} = \gamma_{\min} \geq 1, \\ \gamma_k &= \frac{2}{2 - \varepsilon_k^2} \leq \frac{2}{2 - \varepsilon^2} = \gamma \leq 2, \\ \lambda_k &= \gamma_k \left(\varepsilon_k^2 - \frac{\varepsilon_k^4}{4} \right)^{\frac{1}{2}} \leq \gamma_k \varepsilon_k \leq \gamma \varepsilon. \end{aligned}$$

Bounding the norm terms as $\|\Gamma_I c_I\|_2 \leq \gamma \|c_I\|_2$ and $\|\Lambda_I c_I\|_2 \leq \gamma\varepsilon \|c_I\|_2$ and setting $n_1 = n_2 = n$, $c_S = \min_{i \in I} c_i$, we have that thresholding recovers the generating support I, except with probability $4K^{1-2n}$, whenever

$$c_S \cdot \gamma_{\min} \geq 4\gamma(\bar{\mu} + \varepsilon \nu_Z) \cdot \|c_I\|_2 \sqrt{n \log K}.$$

The final result follows from multiplying both sides by $c_S^{-1} \gamma_{\min}^{-1}$. □

Bibliography

[1] J. Adler. Odl - operator discretization library. `https://github.com/odlgroup/odl`, 2013.

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

[3] M. Aharon, M. Elad, and A.M. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing.*, 54(11):4311–4322, November 2006.

[4] V. Antun, F. Renna, C. Poon, B. Adcock, and A.C. Hansen. On instabilities of deep learning in image reconstruction and the potential costs of AI. *Proceedings of the National Academy of Sciences*, 2020.

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

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

[7] N.G. Behl, C. Gnahm, P. Bachert, M.E. Ladd, and A.M. Nagel. Three-dimensional dictionary-learning reconstruction of 23Na MRI data. *Magnetic Resonance in Medicine*, 75(4):1605–1616, 2016.

[8] G. Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57(297):33–45, March 1962.

[9] J. Caballero, A.N. Price, D. Rueckert, and J.V. Hajnal. Dictionary learning and time sparsity for dynamic MR data reconstruction. *IEEE Transactions on Medical Imaging*, 2014.

[10] E. Candès, L. Demanet, D.L. Donoho, and L. Ying. Fast discrete curvelet transforms. *Multiscale Modeling & Simulation*, 5(3):861–899, 2006.

[11] E. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 52(2):489–509, 2006.

[12] C. Chen, Y. Liu, P. Schniter, N. Jin, J. Craft, O. Simonetti, and R. Ahmad. Sparsity adaptive reconstruction for highly accelerated cardiac MRI. *Magnetic resonance in medicine*, 81(6):3875–3887, 2019.

[13] Y. Chen, X. Ye, and F. Huang. A novel method and fast algorithm for MR image reconstruction with significantly under-sampled data. *Inverse Problems and Imaging*, 4(2):223–240, 2010.

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

[15] I. Daubechies. *Ten Lectures on Wavelets.* CBMS-NSF Lecture Notes. SIAM, 1992.

[16] T.T. Do, L. Gan, N. Nguyen, and T.D. Tran. Sparsity adaptive matching pursuit algorithm for practical compressed sensing. In *2008 42nd Asilomar Conference on Signals, Systems and Computers*, pages 581–587. IEEE, 2008.

[17] J. Dong, W. Wang, W. Dai, M.D. Plumbley, Z. Han, and J. Chambers. Analysis SimCO algorithms for sparse analysis model based dictionary learning. *IEEE Transactions on Signal Processing*, 64(2):417–431, 2016.

[18] D.L. Donoho and M. Elad. Optimally sparse representation in general (non-orthogonal) dictionaries via ℓ_1 minimization. *Proc. Nat. Aca. Sci.,*, 100(5):2197–2202, March 2003.

[19] D.L. Donoho, M. Elad, and V.N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on Information Theory*, 52(1):6–18, January 2006.

[20] D.L. Donoho, Y. Tsaig, I. Drori, and J.L. Starck. Sparse solution of underdetermined systems of linear equations by stagewise orthogonal matching pursuit. *IEEE transactions on Information Theory*, 58(2):1094–1121, 2012.

[21] K. Engan, S.O. Aase, and J.H. Husoy. Method of optimal directions for frame design. In *ICASSP99*, volume 5, pages 2443 – 2446, 1999.

[22] S. Foucart. Hard Thresholding Pursuit: An algorithm for compressive sensing. *SIAM Journal on Numerical Analysis*, 49(6):2543–2563, 2011.

[23] D.A. Freedman. On tail probabilities for martingales. *The Annals of Probability*, 3(1):100–118, 1975.

[24] R. Gribonval, H. Rauhut, K. Schnass, and P. Vandergheynst. Atoms of all channels, unite! Average case analysis of multi-channel sparse recovery using greedy algorithms. *Journal of Fourier Analysis and Applications*, 14(5):655–687, 2008.

[25] D. Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, 2011.

[26] K. Hammernik, T. Klatzer, E. Kobler, M.P. Recht, D.K. Sodickson, T. Pock, and F. Knoll. Learning a variational network for reconstruction of accelerated MRI data. *Magnetic Resonance in Medicine*, 79(6):3055–3071, 2018.

[27] A. Hauptmann, S. Arridge, F. Lucka, V. Muthurangu, and J.A. Steeden. Real-time cardiovascular MR with spatio-temporal artifact suppression using deep learning–proof of concept in congenital heart disease. *Magnetic Resonance in Medicine*, 81(2):1143–1156, 2019.

[28] M.R. Hestenes and E. Stiefel. *Methods of conjugate gradients for solving linear systems*, volume 49. NBS Washington, DC, 1952.

[29] W. Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58:13–30, 1963.

[30] P. Irofti. The effect of atom replacement strategies on dictionary learning. In *iTWIST*, 2016.

[31] A. Kofler, M. Dewey, T. Schaeffter, C. Wald, and C. Kolbitsch. Spatio-temporal deep learning-based undersampling artefact reduction for 2D radial cine MRI with limited training data. *IEEE transactions on medical imaging*, 39(3):703–717, 2019.

[32] A. Kofler, M. Haltmeier, T. Schaeffter, M. Kachelrieß, M. Dewey, C. Wald, and C. Kolbitsch. Neural networks-based regularization for large-scale medical image reconstruction. *Physics in Medicine & Biology*, 65(13):135003, 2020.

[33] K. Kreutz-Delgado, J.F. Murray, B.D. Rao, K. Engan, T. Lee, and T.J. Sejnowski. Dictionary learning algorithms for sparse representation. *Neural Computations*, 15(2):349–396, 2003.

[34] R. Kueng and D. Gross. RIPless compressed sensing from anisotropic measurements. *Linear Algebra and its Applications*, 441:110–123, 2014.

[35] M. Ledoux and M. Talagrand. *Probability in Banach spaces. Isoperimetry and processes.* Springer-Verlag, Berlin, Heidelberg, NewYork, 1991.

[36] M.S. Lewicki and T.J. Sejnowski. Learning overcomplete representations. *Neural Computations*, 12(2):337–365, 2000.

[37] Y. Li, J. Zhang, G. Sun, and D. Lu. The sparsity adaptive reconstruction algorithm based on simulated annealing for compressed sensing. *Journal of Electrical and Computer Engineering*, 2019, 2019.

[38] J.M. Lin. Python Non-Uniform Fast Fourier Transform (PyNUFFT): An accelerated non-Cartesian MRI package on a heterogeneous platform (CPU/GPU). *Journal of Imaging*, 4(3):51, 2018.

[39] Q. Liu, S. Wang, K. Yang, J. Luo, Y. Zhu, and D. Liang. Highly undersampled magnetic resonance image reconstruction using two-level bregman method with dictionary updating. *IEEE Transactions on Medical Imaging*, 32(7):1290–1301, 2013.

[40] M. Lustig, D.L. Donoho, J.M. Santos, and J.M. Pauly. Compressed sensing MRI. *IEEE signal processing magazine*, 25(2):72–82, 2008.

[41] A. Maier, C. Syben, T. Lasser, and C. Riess. A gentle introduction to deep learning in medical image processing. *Zeitschrift für Medizinische Physik*, 29(2):86–101, 2019.

[42] J. Mairal, F. Bach, and J. Ponce. Task-driven dictionary learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):791–804, 2012.

[43] J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, 11:19–60, 2010.

[44] S. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

[45] M.J. Muckley, R. Stern, T. Murrell, and F. Knoll. TorchKbNufft: A high-level, hardware-agnostic non-uniform fast fourier transform. In *ISMRM Workshop on Data Sampling & Image Reconstruction*, 2020.

[46] M.J. Muckley et al. Torch kb-nufft. `https://github.com/mmuckley/torchkbnufft`, 2019.

[47] V. Naumova and K. Schnass. Fast dictionary learning from incomplete data. *EURASIP Journal on Advances in Signal Processing*, 2018(12), 2018.

[48] M.-C. Pali, T. Schaeffter, C. Kolbitsch, and A. Kofler. Adaptive sparsity level and dictionary size estimation for image reconstruction in accelerated 2D radial cine MRI. *Journal of Medical Physics*, 48(1):178–192, 2021.

[49] M.-C. Pali and K. Schnass. Dictionary learning - from local towards global and adaptive. *arXiv:1804.07101v3*, 2021.

[50] 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.

[51] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[52] S. Ravishankar and Y. Bresler. MR image reconstruction from highly undersampled k-space data by dictionary learning. *IEEE Trans. Med. Imag.*, 30(5):1028, 2011.

[53] R. Rubinstein, A. Bruckstein, and M. Elad. Dictionaries for sparse representation modeling. *Proceedings of the IEEE*, 98(6):1045–1057, 2010.

[54] R. Rubinstein, M. Zibulevsky, and M. Elad. Efficient implementation of the K-SVD algorithm using batch orthogonal matching pursuit. Technical report, Computer Science Department, Technion, 2008.

[55] C. Rusu and B. Dumitrescu. Stagewise K-SVD to design efficient dictionaries for sparse representations. *IEEE Signal Processing Letters*, 19(10):631–634, 2012.

[56] J. Schlemper, J. Caballero, J.V. Hajnal, A.N. Price, and D. Rueckert. A deep cascade of convolutional neural networks for dynamic MR image reconstruction. *IEEE Trans. Med. Imag.*, 37(2):491–503, 2018.

[57] K. Schnass. On the identifiability of overcomplete dictionaries via the minimisation principle underlying K-SVD. *Applied and Computational Harmonic Analysis*, 37(3):464–491, 2014.

[58] K. Schnass. Local identification of overcomplete dictionaries. *Journal of Machine Learning Research (arXiv:1401.6354)*, 16(Jun):1211–1242, 2015.

[59] K. Schnass. A personal introduction to theoretical dictionary learning. *Internationale Mathematische Nachrichten*, 228:5–15, 2015.

[60] K. Schnass. Average performance of Orthogonal Matching Pursuit (OMP) for sparse approximation. *IEEE Signal Processing Letters*, 25(12):1865–1869, 2018.

[61] K. Schnass. Convergence radius and sample complexity of ITKM algorithms for dictionary learning. *Applied and Computational Harmonic Analysis*, 45(1):22–58, 2018.

[62] K. Schnass and P. Vandergheynst. Average performance analysis for thresholding. *IEEE Signal Processing Letters*, 14(11):828–831, 2007.

[63] P. Song, L. Weizman, J.F. Mota, Y.C. Eldar, and M.R. Rodrigues. Coupled dictionary learning for multi-contrast MRI reconstruction. *IEEE Transactions on Medical Imaging*, 2019.

[64] J. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

[65] J.A. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, October 2004.

[66] J.A. Tropp. On the conditioning of random subdictionaries. *Applied and Computational Harmonic Analysis*, 25(1-24), 2008.

[67] J. Tsao, P. Boesiger, and K.P. Pruessmann. k-t BLAST and k-t SENSE: Dynamic MRI With High Frame Rate Exploiting Spatiotemporal Correlations. *Magnetic Resonance in Medicine*, 2003.

[68] S. Wang, Y. Chen, T. Xiao, Z. Ke, Q. Liu, and H. Zheng. LANTERN: learn analysis transform network for dynamic magnetic resonance imaging with small dataset. *arXiv preprint arXiv:1908.09140*, 2019.

[69] S. Wang, Z. Ke, H. Cheng, S. Jia, L. Ying, H. Zheng, and D. Liang. DIMENSION: Dynamic MR imaging with both k-space and spatial prior knowledge obtained via multi-supervised network training. *NMR in Biomedicine*, page e4131, 2019.

[70] S. Wang, Q. Liu, Y. Xia, P. Dong, J. Luo, Q. Huang, and D.D. Feng. Dictionary learning based impulse noise removal via l1–l1 minimization. *Signal Processing*, 93(9):2696–2708, 2013.

[71] S. Wang, S. Tan, Y. Gao, Q. Liu, L. Ying, T. Xiao, Y. Liu, X. Liu, H. Zheng, and D. Liang. Learning joint-sparse codes for calibration-free parallel MR imaging. *IEEE transactions on medical imaging*, 37(1):251–261, 2017.

[72] S. Wang, Y. Xia, Q. Liu, P. Dong, D.D. Feng, and J. Luo. Fenchel duality based dictionary learning for restoration of noisy images. *IEEE transactions on image processing*, 22(12):5214–5225, 2013.

[73] Y. Wang, N. Cao, Z. Liu, and Y. Zhang. Real-time dynamic MRI using parallel dictionary learning and dynamic total variation. *Neurocomputing*, 238:410–419, 2017.

[74] Y. Wang and L. Ying. Compressed sensing dynamic cardiac cine MRI using learned spatiotemporal dictionary. *IEEE Transactions on Biomedical Engineering*, 2014.

[75] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 2004.

[76] S. Winkelmann, T. Schaeffter, T. Koehler, H. Eggers, and O. Doessel. An optimal radial profile order based on the golden ratio for time-resolved MRI. *IEEE Transactions on Medical Imaging*, 26(1):68–76, 2006.

[77] B.J. Wintersperger, S.B. Reeder, K. Nikolaou, O. Dietrich, A. Huber, A. Greiser, T. Lanz, M.F. Reiser, and S. O. Schoenberg. Cardiac cine MR imaging with a 32-channel cardiac coil and parallel imaging: impact of acceleration factors on image quality and volumetric accuracy. *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 23(2):222–227, 2006.