Dictionary learning - from local towards global and adaptive

Marie Christine Pali MARIE-CHRISTINE.PALI@UIBK.AC.AT
Karin Schnass KARIN.SCHNASS@UIBK.AC.AT
Department of Mathematics
University of Innsbruck
Technikerstraße 13
6020 Innsbruck, Austria

Abstract

This paper studies the convergence behaviour of dictionary learning via the Iterative Thresholding and K-residual Means (ITKrM) algorithm. On one hand it is proved that ITKrM is a contraction under much more relaxed conditions than previously necessary. On the other hand it is shown that there seem to exist stable fixed points that do not correspond to the generating dictionary, which can be characterised as very coherent. Based on an analysis of the residuals using these bad dictionaries, replacing coherent atoms with carefully designed replacement candidates is proposed. In experiments on synthetic data this outperforms random or no replacement and always leads to full dictionary recovery. Finally the question how to learn dictionaries without knowledge of the correct dictionary size and sparsity level is addressed. Decoupling the replacement strategy of coherent or unused atoms into pruning and adding, and slowly carefully increasing the sparsity level, leads to an adaptive version of ITKrM. In several experiments this adaptive dictionary learning algorithm is shown to recover a generating dictionary from randomly initialised dictionaries of various sizes on synthetic data and to learn meaningful dictionaries on image data.

Keywords: dictionary learning, sparse coding, sparse component analysis, Iterative Thresholding and K-residual Means (ITKrM), replacement, adaptive dictionary learning, parameter estimation.

1. Introduction

The goal of dictionary learning is to decompose a data matrix $Y = (y_1, \dots, y_N)$, where $y_n \in \mathbb{R}^d$, into a dictionary matrix $\Phi = (\varphi_1, \dots, \varphi_K)$, where each column also referred to as atom is normalised, $\|\varphi_k\|_2 = 1$ and a sparse coefficient matrix $X = (x_1, \dots, x_N)$,

$$Y \approx \Phi X \quad \text{and} \quad X \text{ sparse.} \tag{1}$$

The compact data representation provided by a dictionary can be used for data restoration, such as denoising or reconstruction from incomplete information, [12, 27, 29] and data analysis, such as blind source separation, [14, 26, 22, 23]. Due to these applications dictionary learning is of interest to both the signal processing community, where it is also known as sparse coding, and the independent component analysis (ICA) and the blind source separation (BSS) community, where it is also known as sparse component analysis. It also means that there are not only many algorithms to choose from,

[14, 3, 13, 23, 26, 28, 45, 29, 38, 31], but also that theoretical results have started to accumulate, [17, 46, 4, 1, 40, 41, 18, 6, 5, 43, 47, 48, 9, 35]. As our reference list grows more incomplete every day, we point to the surveys [36, 42] as starting points for digging into algorithms and theory, respectively.

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

$$\operatorname*{argmin}_{\Psi \in \mathcal{D}_K, X \in \mathcal{X}_S} \sum_n \|y_n - \Psi x_n\|_2^p. \tag{2}$$

Unfortunately this problem is highly non-convex and as such difficult to solve even in the simplest and most commonly used case $p = 2$. However, randomly initialised alternating projection algorithms, which alternate between (trying to) find the best dictionary Ψ, based on coefficients X, and the best coefficients X, based on a dictionary Ψ, such as K-SVD (K Singular Value Decompositions) for $p = 2$, [3], and ITKrM (Iterative Thresholding and K residual Means) related to $p = 1$, [43], tend to be very successful on synthetic data - usually recovering 90 to 100% of all atoms - and to provide useful dictionaries on image data.

Apart from needing both the sparsity level and the dictionary size as input, the main drawback of these algorithms is that - assuming that the data Y is synthesized from a generating dictionary Φ and randomly drawn S-sparse coefficients X - they have almost no (K-SVD) or comparatively weak (ITKrM) theoretical dictionary recovery guarantees. This is in sharp contrast to more involved algorithms, which - given the correct S, K - have gobal recovery guarantees but due to their computational complexity can at best be used in small toy examples, [4, 2, 6].

There are some interesting exceptions. In [5] Arora et. al. propose an initialisation strategy, that can be used to decide the dictionary size, and several alternating projection algorithms for local refinement, which in combination are proven to recover any well-behaved dictionary. The alternating projection algorithms are very similar in spirit to ITKrM, and we believe that all are our results on the contractive areas of ITKrM can easily be transferred to them. In [47, 48], Sun, Qu and Wright study an algorithm based on gradient descent with a Newton trust region method to escape saddle points and prove recovery if the generating dictionary is a basis. In [35], Qu et. al. study an ℓ^4-norm opimisation programme with spherical constraints for overcomplete dictionary learning. They show that every local minimizer is close to an atom of the target dictionary and that around every saddle point is a large region with negative curvature. These results together with several results in machine learning which prove that non-convex problems can be well behaved, meaning all local minima are global minima, give rise to hope that a similar result can be proved for learning overcomplete dictionaries via alternating projection.

Contribution: In this paper we first study the contractive areas of ITKrM and show that the algorithm contracts towards the generating dictionary under much relaxed conditions compared to those from [43]. We then have a closer a look at experiments where the learned dictionaries do not coincide with the generating dictionary. These spurious dictionaries, which at least experimentally are fixed points, have a very special structure that violates the theoretical conditions for contractivity, that is, they contain two nearly identical atoms.

Unfortunately, these experimental findings indicate that for alternating projection methods not all fixed points correspond to the generating dictionary.
However, based on an analysis of the residuals at dictionaries with the discovered structure, we develop a strategy for finding good candidates to replace coherent atoms. With the help of these replacement candidates, we then tackle one of the most challenging problems in dictionary learning - the automatic choice of the sparsity level S and the dictionary size K. This leads to a version of ITKrM that adapts both the sparsity level and the dictionary size in each iteration. Synthetic experiments show that the resulting algorithm is able recover a generating dictionary without prescribing its size or the sparsity level even in the presence of noise and outliers. Complementary experiments on image data further show that the algorithm learns sensible dictionaries even in practice, where several synthetic assumptions such as homogeneous use of all atoms are unlikely to hold.
Organisation: In the next section we summarise our notational conventions and introduce the sparse signal model on which all our theoretical results are based. In Section 3 we familiarise the reader with the ITKrM algorithm and existing convergence results. We analyse the limitations of existing proofs, develop strategies to overcome them and prove that ITKrM is a contraction towards the generating dictionary on an area much larger than indicated by the convergence radius in [43]. To see whether the non-contractive areas are only an artefact of our proof strategy, we conduct several small experiments. These show that indeed there are fixed points of ITKrM, which are not equivalent to the generating dictionary through reordering and sign flips and violate our conditions for contractivity; in particular, they are very coherent. In Section 4 we analyse the residuals at such bad dictionaries and use those insights to develop a strategy for learning good replacement candidates for coherent atoms. The resulting algorithm is then tested and compared to random replacement on synthetic data.
In Section 5 we then address the big problem how to learn dictionaries without being given the generating sparsity level and dictionary size. This is done by slowly increasing the sparsity level and by decoupling the replacement strategy into separate pruning of the dictionary and adding of promising replacement candidates. Numerical experiments show that the resulting algorithm can indeed recover the generating dictionary from initialisations with various sizes on synthetic data and learn meaningful dictionaries on image data.
In the last section we will sketch how the concepts leading to adaptive ITKrM can be extended to other algorithms such as K-SVD or MOD. Finally, based on a discussion of our results, we will map out future directions of research.

2. Notations and Sparse Signal Model

Before we hit the strings, we will fine tune our notation and introduce some definitions. Usually subscripted letters will denote vectors with the exception of $\varepsilon, \alpha, \omega$, 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 matrix M we denote its (conjugate) transpose by $M^\star$ and its Moore-Penrose pseudo-inverse by $M^\dagger$. 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 $\|M\|_{2,2} \leq \|M\|_F$.
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 [8].
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 $\mathbb{S} = \{1, \dots, S\}$.
We define the **distance** of a dictionary Ψ to a dictionary Φ as

$$d(\Phi, \Psi) := \max_k \min_\ell \|\phi_k \pm \psi_\ell\|_2 = \max_k \min_\ell \sqrt{2 - 2|\langle \phi_k, \psi_\ell \rangle|}. \tag{3}$$

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,

$$d_s(\Phi, \Psi) := \min_{p \in \mathcal{P}} \max_k \|\phi_k \pm \psi_{p(k)}\|_2, \tag{4}$$

where $\mathcal{P}$ is the set of permutations of $\{1, \dots, K\}$. 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

$$\psi_k = \alpha_k \phi_k + \omega_k z_k, \quad \text{for,} \quad \alpha_k := 1 - \varepsilon_k^2/2 \quad \text{and} \quad \omega_k := (\varepsilon_k^2 - \varepsilon_k^4/4)^{\frac{1}{2}}. \tag{5}$$

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 $d(\Psi, \Phi) = \sqrt{2 - 2\alpha_{\min}}$.

2.1 Sparse signal model

As basis for our results we use the following signal model, already used in [40, 41, 43]. Given a $d \times K$ dictionary Φ, we assume that the signals are generated as

$$y = \frac{\Phi x + r}{\sqrt{1 + \|r\|_2^2}}, \tag{6}$$

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 subgaussian vector with parameter ρ, that is, $\mathbb{E}(r) = 0$ and for all vectors v the marginals $\langle v, r \rangle$ are subgaussian 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 $c(1) \geq c(2) \ldots \geq c(K) > S$ 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

$$y = \frac{\Phi x_{c,p,\sigma} + r}{\sqrt{1 + \|r\|_2^2}}. \tag{7}$$

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

$$\frac{c(1)}{c(S)} \leq \gamma_{dyn}, \qquad \frac{c(S+1)}{c(S)} \leq \gamma_{gap}, \qquad \frac{\|c(\mathbb{S}^c)\|_2}{c(1)} \leq \gamma_{app} \qquad \text{and} \qquad \frac{\rho}{c(S)} \leq \gamma_\rho. \tag{8}$$

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.

Algorithm 3.1: ITKrM (one iteration)

Input: Ψ, 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**

    $\bar{\psi}_k \leftarrow \bar{\psi}_k + \left[a_n + P(\psi_k) y_n\right] \cdot \operatorname{sign}(\langle \psi_k, y_n \rangle)$; // atom update

  end

end

$\Psi \leftarrow \left(\bar{\psi}_1/\|\bar{\psi}_1\|_2, \ldots, \bar{\psi}_K/\|\bar{\psi}_K\|_2\right)$; // atom normalisation

Output: Ψ

Apart from these worst case bounds we will also use three other signal statistics,

$$\gamma_{1,S} := \mathbb{E}_c\left(\|c(\mathbb{S})\|_1\right), \qquad \gamma_{2,S} := \mathbb{E}_c\left(\|c(\mathbb{S})\|_2^2\right), \qquad C_r := \mathbb{E}_r\left(\frac{1}{\sqrt{1+\|r\|_2^2}}\right). \tag{9}$$

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

$$C_r \geq \frac{1 - e^{-d}}{\sqrt{1 + 5d\rho^2}}, \tag{10}$$

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 [41] 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. Global behaviour patterns of ITKrM

The iterative thresholding and K residual means algorithm (ITKrM) was introduced in [43] as modification of its much simpler predecessor ITKsM, which uses signal means instead of residual means. As can be seen from the summary in Algorithm 3.1 each signal can be processed and discarded, thus making the algorithm suitable for a sequential version and parallelisation. The determining factors for the computational complexity 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 $O(S^2 dN)$, corresponding to the QR decompositions of $\Psi_{I_n^t}$. However, since usually the achievable precision in the learning is limited by the number of

available training signals rather than the numerical precision, it is computationally more efficient to precompute the Gram matrix $\Psi^\star\Psi$ and calculate the projections less stably via the eigenvalue decompositions of $\Psi^\star_{I_n^t}\Psi_{I_n^t}$, corresponding to an overall cost $O(S^3N)$. Another good property of the ITKrM algorithm is that it is proven to converge locally to a generating dictionary. This means that if the data is homogeneously S-sparse in a dictionary Φ, where $S \lesssim \mu^{-2}$, and we initialise with a dictionary Ψ within radius $O(1/\sqrt{S})$, $d(\Psi, \Phi) \lesssim 1/\sqrt{S}$, then ITKrM using $N = O(K \log K)$ samples in each iteration will converge to the generating dictionary, [43]. In simulations on synthetic data ITKrM 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.
Last but not least, ITKrM is not just a pretty toy with theoretical guarantees but on image data produces dictionaries of the same quality as K-SVD in a fraction of the time, [30].
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 cousin ITKsM, which when initialised randomly performs much worse both on synthetic and image data, we can prove a convergence radius of size $O(1/\sqrt{\log K})$. Therefore, in the next section 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.

3.1 Contractive areas of ITKrM

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

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

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

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

The term S_1 captures the errors thresholding makes in estimating the supports I_n and signs $\sigma_n(k)$ when using the current estimate Ψ. It is (sufficiently) small as long as $d(\Phi, \Psi) \lesssim 1/\sqrt{\log K}$. The second term S_2 captures the difference between the residual using the

current estimate and the true dictionary, 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, so as long as the number of signals N is large enough, the last term will concentrate arbitrarily close to ϕ_k.
As we can see, 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 $d(\Phi, \Psi) \approx 1/\sqrt{S}$, 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, if the current dictionary estimate Ψ is itself a well-conditioned and incoherent matrix, results on the conditioning of random subdictionaries, [50, 10], tell us that 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 that we can check 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 our previous estimates we bounded the error per atom by the probability of thresholding failing multiplied with the norm bound on the difference of the projections. While simple, this strategy 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 was 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. To keep the flow of the paper we will state it in an informal version and refer the reader to Appendix A for the exact statement and its proof.

Theorem 1 *Assume that the sparsity level of the training signals scales as $S \lesssim \mu(\Phi)^{-2}/\log K$ and that the number of training signals scales as $N \approx SK \log K$. Further, assume that the coherence and operator norm of the current dictionary estimate Ψ satisfy,*

$$\mu(\Psi) \lesssim \frac{1}{\log K} \quad \text{and} \quad \|\Psi\|_{2,2}^2 \lesssim \frac{K}{S \log K}. \tag{11}$$

If the distance of Ψ to the generating dictionary Φ satisfies either

a) $\frac{1}{\sqrt{S}} \lesssim d(\Psi, \Phi) \lesssim \frac{1}{\sqrt{\log K}}$ *or*

b) $d(\Psi, \Phi) \gtrsim \frac{1}{\sqrt{\log K}}$ *but the cross-Gram matrix $\Phi^\star \Psi$ is diagonally dominant in the sense that*

$$\min_k |\langle \phi_k, \psi_k \rangle| \gtrsim \log K \cdot \max \left\{ \mu(\Phi, \Psi), \|\Phi\|_{2,2} \sqrt{S/(K \log K)} \right\}, \tag{12}$$

then one iteration of ITKrM will reduce the distance by at least a factor $\kappa < 1$, meaning

$$d(\bar{\Psi}, \Phi) < \kappa \cdot d(\Psi, \Phi).$$

The first part of the theorem simply says that, excluding dictionaries Ψ that are coherent or have large operator norm, ITKrM is a contraction on a ball of radius $1/\sqrt{\log K}$ around the generating dictionary Φ. To better understand the second part of the theorem, we have a closer look at the conditions on the cross-Gram matrix $\Phi^\star\Psi$ in (12). The fact that the diagonal entries have to be larger than $\|\Phi\|_{2,2}\sqrt{(S\log K)/K}$ 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 a well-conditioned dictionary, satisfying $\|\Phi\|_{2,2}^2 \approx K/d$, this means that

$$d(\Phi,\Psi) \lesssim \left(2 - 2\sqrt{\frac{S\log K}{d}}\right)^{1/2}. \tag{13}$$

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

$$\min_k |\langle \phi_k, \psi_k\rangle| \gtrsim \log K \cdot \max_{j\neq k} |\langle \phi_k, \psi_j\rangle|. \tag{14}$$

This condition captures our intuition that two estimated atoms should not point to the same generating atom and provides a bound for sufficient separation.
One thing that has to be noted about the result above is that it does not guarantee convergence of ITKrM since it is only valid for one iteration. To prove convergence of ITKrM, we need to additionally prove that $\bar{\Psi}$ inherits from Ψ the properties that are required for being a contraction, which is part of our future goals. Still, the result goes a long way towards 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 $\|\Phi\|_{2,2}^2 \approx K/d$. If $\psi_k = \alpha_k\phi_k + \omega_k z_k$, where the perturbation vectors z_k are drawn uniformly at random from the unit sphere orthogonal to ϕ_k, then with high probability for all $j \neq k$ we have

$$|\langle \phi_k, z_j\rangle| \lesssim \sqrt{\log K/d} \quad \text{and} \quad |\langle z_k, z_j\rangle| \lesssim \sqrt{\log K/d} \tag{15}$$

and consequently for all possible α_k

$$\mu(\Psi) \lesssim \sqrt{4\log K/d} \quad \text{and} \quad \mu(\Phi,\Psi) \lesssim \sqrt{2\log K/d}. \tag{16}$$

Also with high probability the operator norm of the matrix $Z = (z_1, \dots z_K)$ is bounded by $\|Z\|_{2,2} \lesssim \sqrt{\log K}$, [49], 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

$$\alpha_{\min} \gtrsim \sqrt{\frac{S(\log K)^2}{d}} \qquad \Leftrightarrow \qquad d(\Phi,\Psi) \lesssim \left(2 - 2\sqrt{\frac{S(\log K)^2}{d}}\right)^{1/2}. \tag{17}$$

A fully random initialisation will have small coherence and operator norm with high probability. While it is also exponentially more likely to satisfy the cross-coherence property

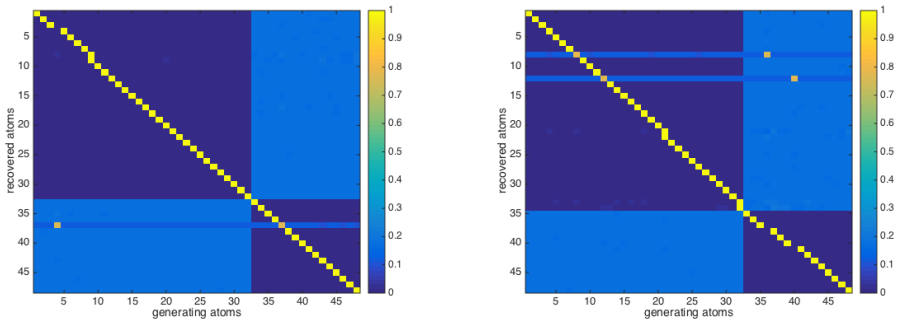

Figure 1: Cross-Gram matrices $\Psi^\star\Phi$ for recovered dictionaries with 2 (left) and 4 (right) missing atoms.

than to be within distance $1/\sqrt{S}$ or $1/\sqrt{\log K}$ to the generating dictionary, the absolute probability of having the cross-coherence property is still very small. This leads to the question whether in practice the cross-coherence property is actually necessary for convergence. Experiments in the next section will provide evidence that it is practically relevant, in the sense, that whenever ITKrM does not recover the generating dictionary, it produces a dictionary not satisfying the cross-coherence property.

3.2 Bad dictionaries

From [43] 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[1]. 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 (7), where the coefficient sequences c are constructed by chosing $b \in [0.9, 1]$ 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. 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 3.2 - 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.
So in the most simple case of 2 missing atoms (after rearranging and sign flipping the atoms

1. All experiments and resulting figures can be reproduced using the matlab toolbox available at `https://www.uibk.ac.at/mathematik/personal/schnass/code/adl.zip`

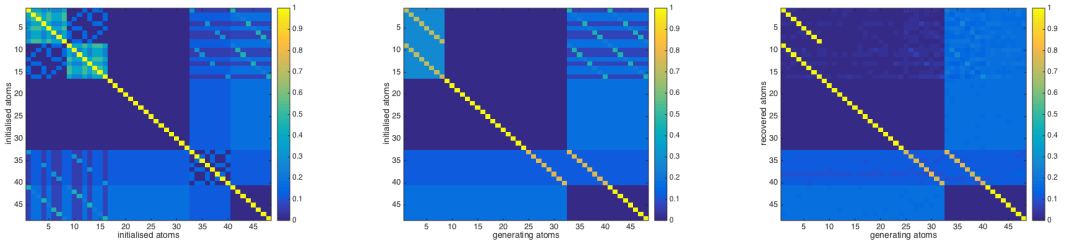

Figure 2: Absolute values of the Gram matrix of a bad initial dictionary $\Psi^\star\Psi$ (left), its cross-Gram matrix with the generating dictionary $\Psi^\star\Phi$ (middle) and the cross-Gram matrix of the recovered dictionary after 25 iterations of ITKrM $\bar{\Psi}^\star\Phi$ (right).

in Φ) the recovered (and rearranged) dictionary Ψ has the form

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

Looking back to our characterisation of the contractive areas in the last section, we see that such a dictionary or a slightly perturbed version of it clearly cannot have the necessary cross-coherence property with any reasonably incoherent dictionary. A complete proof showing that Ψ is indeed a stable spurious fixed point is unfortunately too long to be included here but in preparation. In the meantime we refer the interested reader to the preprint version where a sketch of the proof can be found, [44]. We also provide some intuition why dictionaries of the above form are stable in the next section.

Here we just want to add, that while a dictionary with coherence $\mu \approx 1$ clearly does not satisfy the conditions for contractivity, the reverse is not true. On the contrary two estimated atoms pointing to the same generating atom ϕ_j can be very incoherent even if they are both already quite close to ϕ_j. For instance, if $\psi_{j_\pm} \approx \alpha_j \phi_j \pm \omega_j z_j$ where z_j is a balanced sum of the other atoms $z_j \approx \sum_{i \neq j} \phi_i \sigma(i)$, we have $|\langle \psi_{j_+}, \psi_{j_-} \rangle| = \alpha_j^2 - \omega_j^2$, meaning approximate orthogonality at $\alpha_j = 1/\sqrt{2}$. Using these ideas we can construct well-conditioned and incoherent initial dictionaries Ψ, with abritrary distances $d(\Psi, \Phi) \gtrsim 1/\sqrt{2}$ to the generating dictionary, so that things will go maximally wrong, meaning we end up with a lot of double and 1:1 atoms. Figure 3.2 shows an example of a bad initialisation with coherence $\mu = 0.52$, leading to 16 missing atoms. The accompanying matlab toolbox provides more examples of these bad initialisations to observe convergence, to play around with and to inspire more evil constructions.

Summarising the two last subsections we see that ITKrM may not be a contraction if the current dictionary estimate is too coherent, has large operator norm or if two atoms are close to one generating atom. Both coherence and operator norm of the estimate could be calculated after each iteration to check whether ITKrM is going in a good direction. Unfortunately, the diagonal dominance of the cross-Gram matrix, which prevents two estimated atoms to be close to the same generating atom, cannot be verified immediately. However, the most likely outcome of this situation is that both these estimated atoms converge to the same generating atom, meaning that eventually the estimated dictionary will be coher-

ent. This suggests that in order to improve the global convergence behaviour of ITKrM, we should control the coherence of the estimated dictionaries. One strategy to incorporate incoherence into ITKrM could be adding a penalty for coherent dictionaries. The main disadvantages of this strategy, apart from the fact that ITKrM is not straightforwardly associated to an optimisation programme, are the computational cost and the fact that penalties tend to complicate the high-dimensional landscape of basins of attractions which further complicates convergence. Therefore, we will use a different strategy which allows us to keep the high percentage of correctly recovered atoms and even use the information they provide for identifying the missing ones: replacement.

4. Replacement

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 [3], employ additionally in each iteration. While randomly drawing a replacement 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 and what to do instead, we first have a look at the shape of the signal residuals at one of the bad dictionaries, identified in the last section.

4.1 Learning from bad dictionaries

We start with an analysis of thresholding in case the current dictionary Ψ contains one double atom, $\psi_1 = \psi_2 = \phi_1$, and one 1:1 atom, $\psi_K \propto \phi_2 + \phi_K$; for the other atoms we have $\psi_k = \phi_k$. We will also keep track of what would happen if we replaced one of the double atoms with a vector drawn uniformly at random from the unit sphere, which we label ψ_0. For simplicity we assume that the signals follow the sparse model in (7) with constant coefficients $c_i = 1$ for $i \leq S$ and $c_i = 0$ for $i \geq 0$ and no noise, and that $\langle \phi_2, \phi_K \rangle \geq 0$. We also adopt the notation $I_{\ell \leftrightarrow k} := (I \setminus \{\ell\}) \cup \{k\}$. Note that we have

$$\begin{aligned} |\langle \psi_k, \phi_k \rangle| &= 1 \quad \text{for} \quad k \notin \{2, K\} \\ \text{and} \quad |\langle \psi_k, \phi_i \rangle| &\leq \mu \quad \text{for} \quad k \neq K,\ i \notin \{1, 2, k\}. \end{aligned}$$

We also have $|\langle \psi_2, \phi_1 \rangle| = 1$ as well as

$$\begin{aligned} |\langle \psi_K, \phi_2 \rangle| = |\langle \psi_K, \phi_K \rangle| &= \frac{|\langle \phi_2 + \phi_K, \phi_K \rangle|}{\sqrt{2 + 2\langle \phi_2, \phi_K \rangle}} = \sqrt{\frac{1 + \langle \phi_2, \phi_K \rangle}{2}} \geq \frac{1}{\sqrt{2}} \\ \text{and} \quad |\langle \psi_K, \phi_i \rangle| &= \frac{|\langle \phi_2 + \phi_K, \phi_i \rangle|}{\sqrt{2 + 2\langle \phi_2, \phi_K \rangle}} \leq \frac{2\mu}{\sqrt{2}} \leq \sqrt{2}\mu \quad \text{for} \quad i \notin \{2, K\} \end{aligned}$$

Since ψ_0 is drawn uniformly at random from the d-dimensional unit sphere, we have for any fixed vector v that

$$\mathbb{P}(|\langle \psi_0, v \rangle| \geq t) \leq 2 \exp\left(-\frac{t^2 d}{2}\right).$$

This means that with very high probability $|\langle\psi_0, \phi_k\rangle| \lesssim \sqrt{\log K/d}$ for all k.
If we draw a random support I of size S, (a random permutation), then with probability $\binom{K-3}{S}/\binom{K}{S} \approx \left(1 - \frac{S}{K}\right)^3$ it does not contain $1, 2, K$, meaning $I \cap \{1, 2, K\} = \emptyset$. We then have

$$\begin{aligned}
k \in I: \quad & |\langle\psi_k, y\rangle| = |\langle\psi_k, \phi_k\rangle\sigma_k c_k + \sum_{i\in I, i\neq k} \langle\psi_k, \phi_i\rangle\sigma_i c_i| \geq 1 - (S-1)\mu, \\
k \in I^c \setminus \{0, K\}: \quad & |\langle\psi_k, y\rangle| = |\sum_{i\in I} \langle\psi_k, \phi_i\rangle\sigma_i c_i| \leq S\mu, \\
k = K: \quad & |\langle\psi_k, y\rangle| = |\sum_{i\in I} \langle\psi_K, \phi_i\rangle\sigma_i c_i| \leq \sqrt{2}S\mu \\
k = 0: \quad & |\langle\psi_k, y\rangle| = |\sum_{i\in I} \langle\psi_0, \phi_i\rangle\sigma_i c_i| \leq S\sqrt{\log K/d}.
\end{aligned}$$

So no matter whether the dictionary contains a double atom or a random replacement atom, thresholding will correctly identify the support, $I^t = I$, and the residual will be zero

$$a = y - P(\Psi_{I^t})y = \Phi_I x_I - P(\Phi_I)\Phi_I x_I = 0.$$

Next we have a look at supports containing 1 but not 2 or K, meaning $I \cap \{1, 2, K\} = \{1\}$. Such a support is drawn with probability $\binom{K-3}{S-1}/\binom{K}{S} \approx \frac{S}{K}\left(1 - \frac{S}{K}\right)^2$. As before we get the following bounds

$$\begin{aligned}
k \in I \cup \{2\}: \quad & |\langle\psi_k, y\rangle| \geq 1 - (S-1)\mu, \\
k \in I^c \setminus \{0, 2, K\}: \quad & |\langle\psi_k, y\rangle| \leq S\mu,
\end{aligned}$$

as well as $|\langle\psi_K, y\rangle| \leq \sqrt{2}S\mu$ and $|\langle\psi_0, y\rangle| \leq S\sqrt{\log K/d}$, which ensures that $I^t \subseteq I \cup \{2\}$. If we are lucky and $|\langle\psi_2, y\rangle| = |\langle\psi_1, y\rangle| = \min_{i\in I} |\langle\psi_i, y\rangle|$ or in the case of the dictionary with the random replacement atom, thresholding recovers $I^t = I$ or $I^t = I_{1\leftrightarrow 2} \equiv I$ and the residual is again zero. If we are less lucky, we miss one of the relevant atoms indexed by $i \in I$, meaning $I^t = I_{i\leftrightarrow 2}$. In this case the residual will be close to ϕ_i,

$$a = y - P(\Psi_{I^t})y = \Phi_I x_I - P(\Phi_{I\setminus\{i\}})\Phi_I x_I = x_i(\phi_i - P(\Phi_{I\setminus\{i\}})\phi_i) \approx \pm\phi_i,$$

since $\|P(\Phi_{I\setminus\{i\}})\phi_i\|_2 \leq \mu\sqrt{S/(1-S\mu)}$. Note that for any fixed $i \notin \{1, 2, K\}$, the probability that both $\{1, i\} \subseteq I$ is bounded by $\binom{K-3}{S-2}/\binom{K}{S} \approx \frac{S^2}{K^2}\left(1 - \frac{S}{K}\right)$, so a residual close to ϕ_i appears with probability less than $\frac{S^2}{K^2}$.
Finally, we analyse what happens when the support contains K but not 1 or 2, meaning $I \cap \{1, 2, K\} = \{K\}$. Again this occurs with probability $\approx \frac{S}{K}\left(1 - \frac{S}{K}\right)^2$. We then have

$$\begin{aligned}
k \in I \setminus \{K\}: \quad & |\langle\psi_k, y\rangle| = |\langle\psi_k, \phi_k\rangle\sigma_k c_k + \sum_{i\in I, i\neq k} \langle\psi_k, \phi_i\rangle\sigma_i c_i| \geq 1 - (S-1)\mu, \\
k = K: \quad & |\langle\psi_k, y\rangle| = |\langle\psi_K, \phi_K\rangle\sigma_K c_K + \sum_{i\in I, i\neq K} \langle\psi_K, \phi_i\rangle\sigma_i c_i| \geq \frac{1 - 2(S-1)\mu}{\sqrt{2}},
\end{aligned}$$

as well as $|\langle\psi_0, y\rangle| \leq S\sqrt{\log K/d}$ and $|\langle\psi_k, y\rangle| \leq S\mu$ for all other atoms, which shows that for both types of dictionary Ψ (with double or random replacement atom) thresholding will recover $I^t = I$. For the residual we get

$$\begin{aligned} a = y - P(\Psi_{I^t})y = [\mathbb{I}_d - P(\Psi_I)]\Phi_I x_I = x_K[\mathbb{I}_d - P(\Psi_I)]\phi_K \\ = x_K[\mathbb{I}_d - P(\Psi_I)][P(\psi_K) + Q(\psi_K)]\phi_K \\ = x_K[\mathbb{I}_d - P(\Psi_I)]Q(\psi_K)\phi_K \\ \approx \pm\tfrac{1}{2}(\phi_K - \phi_2), \end{aligned}$$

where we have used that $Q(\psi_K)\phi_K = \phi_K - \langle\psi_K, \phi_K\rangle\psi_K = \frac{1}{2}(\phi_K - \phi_2)$ and

$$\|P(\Psi_I)Q(\psi_K)\phi_K\|_2 \leq \tfrac{1}{2} \cdot \|\Psi_I^\dagger\|_{2,2} \cdot \|\Psi_I^\star(\phi_K - \phi_2)\|_2 \leq \mu \cdot \sqrt{S/(1-2S\mu)}.$$

The analysis of the case $I \cap \{1, 2, K\} = \{2\}$ is analogue to the one above and shows that the residual again is close to $\pm\frac{1}{2}(\phi_K - \phi_2)$. Summarising our analysis so far, we see that with probability $\left(1 - \frac{S}{K}\right)^3$ the residual will be zero (or close to zero in the noisy case), with probability at most $\frac{S^2}{K^2}$ it will be close to ϕ_i for each $i \notin \{1, 2, K\}$ and with probability $\frac{2S}{K}\left(1 - \frac{S}{K}\right)^2$ it will be close to a scaled version of

$$\psi_{K'} = \frac{\phi_K - \phi_2}{\sqrt{2 - 2\langle\phi_2, \phi_K\rangle}}.$$

Also after covering all supports except those where $|I \cap \{1, 2, K\}| \geq 2$, which together have probability $\approx \frac{3S^2}{K^2}$, we have not encountered a situation, where the randomly chosen atom ψ_0 would have been picked. Indeed a more detailed analysis to be found in [32] shows that ψ_0 only has a chance to be picked if $I \cap \{1, 2, K\} = \{2, K\}$. Moreover, for $\sigma_2 = \sigma_K$ we have $a = y - P(\Psi_{I\setminus\{2\}})y = 0$. So even if ψ_0 is picked, it will not be pulled in a useful direction. Indeed ψ_0 will only be picked and pulled in a good direction if $\sigma_2 = -\sigma_K$ and therefore

$$a = y - P(\Psi_{I^t})y \approx y - P(\Psi_{I\setminus\{2,K\}})y \approx \alpha \cdot \psi_{K'},$$

for some scaling factor $\alpha = \pm\sqrt{2 - 2\langle\phi_2, \phi_K\rangle}$. This case is also the only case, where we have the chance of accidentally picking ψ_1 or ψ_2 and having them distorted in the direction of $\psi_{K'}$. Note that in case $I \cap \{1, 2, K\} = \{1, 2\}$ or $\{1, K\}$ we recover $I^t = I$ or $I^t = I_{K\leftrightarrow 2}$ and so $a \approx \pm\phi_2$ or $a \approx \pm\phi_K$, but due to the random signs the pull in these useful directions cancels out, and similarly for $\{1, 2, K\} \subseteq I$. The intuition why configurations like Ψ are stable is that this case is so rare that ψ_1 resp. ψ_2 cannot be sufficiently perturbed to change the behaviour of thresholding in the next iteration.
So rather than hoping for the at best unlikely distortion of ψ_0 or $\psi_{1/2}$ towards $\psi_{K'}$ we will use the fact that most non-zero residuals (or residuals not just consisting of noise) are close to scaled versions of $\psi_{K'}$ and recover $\psi_{K'}$ directly. Indeed a more general analysis, to be found in [32], shows that for dictionaries containing several double atoms and 1:1 combinations, meaning after rearranging and resigning the dictionaries Φ, Ψ we have $\psi_k = \phi_k$ for $k > 3L$ as well as

$$\psi_\ell = \psi_{L+\ell} = \phi_\ell \quad \text{and} \quad \psi_{2L+\ell} = \frac{\phi_\ell + \phi_{2L+\ell}}{\sqrt{2 + 2\langle\phi_\ell, \phi_{2L+\ell}\rangle}} \quad \text{for} \quad \ell = 1 \dots L,$$

the residuals are 1-sparse in the L complementary 1:1 combinations.

$$\psi_\ell^c := \frac{\phi_\ell - \phi_{2L+\ell}}{\sqrt{2 - 2\langle \phi_\ell, \phi_{2L+\ell}\rangle}} \quad \text{for} \quad \ell = 1 \dots L.$$

This suggests to use ITKrM with sparsity level 1, which reduces to ITKsM or line clustering, on the residuals to directly recover the 1:1 complements as replacement candidates. Replacing the double atom ψ_ℓ by ψ_ℓ^c, in the next iteration it will be serious competition for $\psi_{2L+\ell}$ in the thresholding of all signals containing either ϕ_ℓ or $\phi_{2L+\ell}$. This iteration will then create a first imbalance of the ratio between ϕ_ℓ and $\phi_{2L+\ell}$ within one or both of the estimated atoms, making one the more likely choice for ϕ_ℓ and the other the more likely choice for $\phi_{2L+\ell}$ in the subsequent iteration. There the imbalance will be further increased until a few iterations later we finally have $\psi_\ell \approx \phi_\ell$ and $\psi_{2L+\ell} \approx \phi_{2L+\ell}$ or the other way around.
We can also immediately see the advantages this ITKsM/clustering approach provides over other residual based replacement strategies, such as using the largest residual or using the largest principal components or [38, 21]. 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 most likely be a linear combination of several 1:1 complementary atoms and as such less serious competition for the original 1:1 combinations during thresholding. Additionally to lower chances of being picked, they will also need more iterations to determine which one will rotate into which place.
After learning enough from bad dictionaries to inspire a promising replacement strategy, the next subsection will deal with its practical implementation.

4.2 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 be mature yet, meaning they might not have converged yet.

Efficient learning of replacement atoms.
To solve this problem, observe that the number of replacement candidates, stored in $\Gamma = (\gamma_1, \dots \gamma_L)$, 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 we replace? One strategy that has been employed for instance in the context of analysis operator learning, [11], is to average the two atoms, that is to set $\psi_k^{new} = \psi_k + \operatorname{sign}(\langle \psi_k, \psi_{k'}\rangle)\psi_{k'}$. 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 $\psi_k^{new} = v(k)\psi_k + \operatorname{sign}(\langle\psi_k, \psi_{k'}\rangle)v(k')\psi_{k'}$. 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, $\max_k |\langle\gamma_\ell, \phi_k\rangle| \geq \mu_{\max}$. Note that in a perfectly S-sparse setting this is not very likely since the residuals we are summing up contain mainly noise or missing 1:1 complements and therefore add up to noise or the desired 1:1 complements. However it might be a problem if we underestimate the sparsity level in the learning. If we use $\tilde{S} < S$, the residuals are still at least $S - \tilde{S}$ sparse in the dictionary, so some of our replacement candidates might be near copies of already recovered atoms in the dictionary. 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

$$\mathbb{P}(|\langle\gamma_\ell, r\rangle| \geq \tau\|r\|_2) \leq 2\exp\left(-\frac{d\tau^2}{2}\right). \tag{18}$$

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.
The threshold should also help in the earlier mentioned case of underestimating the sparsity level. There one could imagine the candidates to be poolings of already recovered atoms,

that is, $\gamma_\ell \approx \sum_{j\in J_\ell} \pm\phi_j/\sqrt{|J_\ell|}$, which are sufficiently incoherent to the dictionary atoms to pass the coherence test. If the residuals are homogenously $S-\tilde{S}$ sparse in the original dictionary, the candidate atom γ_ℓ will be picked if ϕ_j approximates the residual best for a $j \in J_\ell$. If additionally the sets J_ℓ are disjoint, atoms corresponding to a bigger atom pool are (up to a degree) more likely to be choosen than those corresponding to a smaller pool. The threshold helps favour candidates associated to small pools, which have bigger inner products, since $|\langle \gamma_\ell, a_n\rangle| \approx 1/\sqrt{S|J_\ell|}$. This is desirable since the candidate closest to the missing atom will correspond to a smaller pool. After all, a candidate containing in its pool the missing atom (1:1 complement) will be soon distorted towards this atom since the sparse residual coefficient of the missing atom will be on average larger than those of the other atoms, thus reducing the effective size of the pool.

Dealing with unused atoms.

Before implementing our new replacement strategy, let us address another less frequently activated safeguard included in most dictionary learning algorithms: the handling of dictionary atoms that are never selected and therefore have a zero update. As in the case of coherent atoms the standard procedure is replacement of such an atom with a random redraw, which however comes with the problems discussed above. Fortunately our replacement candidates again provide an efficient alternative. 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 the above considerations leads to the augmented ITKrM algorithm, which is summarised in Algorithm C.1 while the actual procedure for replacing coherent atoms is described in Algorithm C.2, both to be found in Appendix C. With these details fixed, the next step is to see how much the invested effort will improve dictionary recovery.

4.3 Numerical Simulations

In this subsection we will verify that replacing coherent atoms improves dictionary recovery and test whether our strategy improves over random replacement. Our main setup is the following:

Generating dictionary: As generating dictionary Φ we use a dictionary of size $K = 192$ in $\mathbb{R}^d$ with $d = 128$, where the atoms are drawn i.i.d. from the unit sphere.

(Sparse) training signals: We generate S-sparse training signals according to our signal model in (7) as

$$y = \frac{\Phi x_{c,p,\sigma} + r}{\sqrt{1+\|r\|_2^2}}. \tag{19}$$

For every signal a new sequence c is generated by drawing a decay factor q uniformly at random in $[0.9, 1]$ and setting $c_i = c_q q^{i-1}$ for $i \leq S$ and 0 else, where $c_q := \frac{1-q}{1-q^S}$ so that $\|c\|_2 = 1$. The noise is centered Gaussian noise with variance $\rho^2 = (16d)^{-1}$, leading to a signal to noise ratio of SNR = 16. We will consider two types of training signals. The first type consists of 6-sparse signals with 5% outliers, that is, we randomly select 5% of the sparse signals and replace them with pure Gaussian noise of variance $1/d^2$. The second type consists of 25% 4-sparse signals, 50% 6-sparse signals and 25% 8-sparse signals, where

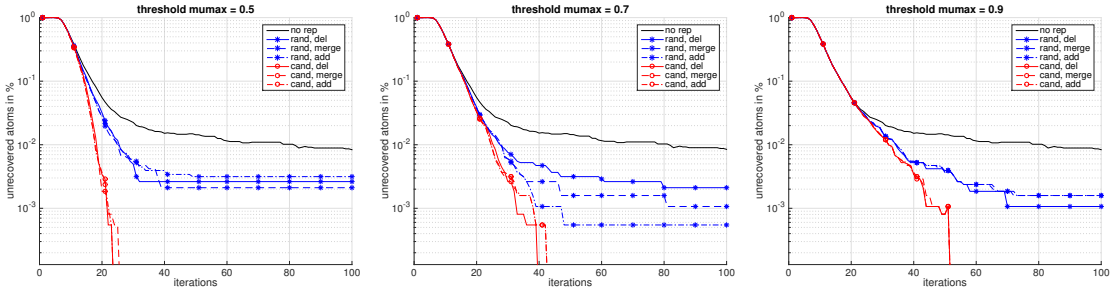

Figure 3: Recovery rates of ITKrM without replacement, random and candidate replacement for various coherence thresholds $\mu_{\max}$ and atom combination strategies.

again 5% are replaced with pure Gaussian noise. In each iteration of ITKrM we use a fresh batch of $N = 120000$ training signals. Unless specified otherwise, the sparsity level given to the algorithm is $S_e = 6$.

Replacement candidates: During every iteration of ITKrM we learn $L = \lfloor \log d \rceil = 5$ replacement candidates using $m = \lfloor \log d \rceil = 5$ iterations each with $N_\Gamma = \lfloor N/m \rfloor$ signals.

Initialisations: The dictionary Ψ containing K atoms as well as the replacement candidates are initialised by drawing vectors i.i.d. from the unit sphere. In case of random replacement we use the initialisations of the replacement candidates. All our results are averaged over 20 different initialisations.

Replacement thresholds: We will compare the dictionary recovery for various coherence thresholds $\mu_{\max} \in \{0.5, 0.7, 0.9\}$, and all three combination strategies, adding, deleting and merging. We also employ an additional safeguard and replace atoms, which have not been used at all or which have energy smaller than 0.001 before normalisation, if after replacement of coherent atoms we have candidate atoms left.

Recovery threshold: We use the convention that a dictionary atom ϕ_k is recovered if $\max_j |\langle \phi_k, \psi_j \rangle| \geq 0.99$.

The results of our first experiment[2], which explores the efficiency of replacement using our candidate strategy in comparison to random or no replacement on 6-sparse signals as described above, are depicted in Figure 3. We can see that for all three considered coherence thresholds $\mu_{\max} \in \{0.5, 0.7, 0.9\}$, our replacement strategy improves over random or no replacement. So while after 100 iterations ITKrM without replacement misses about 1% of the atoms and with random replacement about 0.1%, it always finds the full dictionary after at worst 55 iterations using the candidate atoms. Contrary to random replacement the candidate based strategy also does not seem sensitive to the combination method. Another observation is that candidate replacement leads to faster recovery the lower the coherence threshold is, while the average performance for random replacement is slightly better for the higher thresholds. This is connected to the average number of replaced atoms in each run, which is around 16 for $\mu_{\max} = 0.5$, around 3.8 for $\mu_{\max} = 0.7$ and around 0.8 for $\mu_{\max} = 0.9$, since for the candidate replacement there is no risk of replacing a coherent atom that might

2. As already mentioned, all experiments can be reproduced using the matlab toolbox available at `https://www.uibk.ac.at/mathematik/personal/schnass/code/adl.zip`

still change course and converge to a missing generating atom with something useless. For the sake of completeness, we also mention that in none of the trials replacement of unused atoms is ever activated.

In our second experiment we explore the performance of candidate replacement for the more interesting (realistic) type of signals with varying sparsity levels. Since the signals can be considered 4, 6 or 8 sparse we compare the performance of ITKrM using all three possibilities, $S_e \in \{4, 6, 8\}$ and a fixed replacement threshold $\mu_{\max} = 0.7$. The results are shown in Figure 4. As before, candidate replacement outperforms random or no replacement and leads to 100% recovery in all cases. Comparing the speed of convergence we see that it is higher the lower the sparsity level is, so for $S_e = 4$ we get 100% recovery after about 30 iterations, for $S_e = 6$ after about 65 iterations and for $S_e = 8$ after 75 iterations. This would suggest that for the best performance we should always pick a lower than average sparsity level. However, the speed of convergence for $S_e = 4$ comes at the price of precision, as can be seen in the small table below, which lists both the average distance $d(\Psi, \Phi)$ of the recovered dictionaries from the generating dictionary after 100 iterations as well as the mean atom distances $d_1(\Psi, \Phi) := \frac{1}{K} \sum_k \|\phi_k - \psi_k\|_2$. For both distances there is an increase

	$S_e = 4$	$S_e = 6$	$S_e = 8$
$d(\Psi, \Phi)$	0.0392	0.0312	0.0304
$d_1(\Psi, \Phi)$	0.0322	0.0256	0.0250

in precision, going from $S_e = 4$ to $S_e = 6$, but hardly any improvement by going from $S_e = 6$ to $S_e = 8$. This suggests to choose the average or a slightly higher than average sparsity level. Alternatively, to get the best of both worlds, one should start with a smaller sparsity level and then slowly increase to the average sparsity level. Unfortunately this approach relies on the knowledge of the average sparsity level, which in practice is unknown. Considering that also the size of the dictionary is unknown this can be considered a minor problem. After all, if we underestimate the dictionary size, this will limit the final precision more severely. Assume for instance that we set $K - 1$ instead of K. In this case recovering a dictionary Ψ with $K - 2$ generating atoms plus one 1:1 combination of two generating

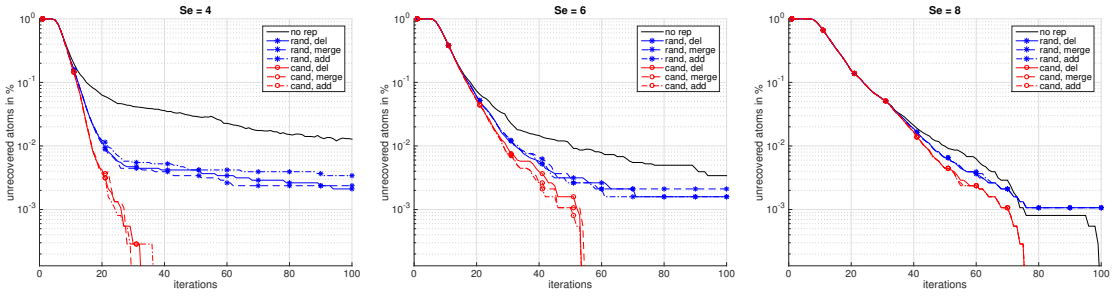

Figure 4: Recovery rates of ITKrM without replacement, random and candidate replacement for various input sparsity levels S_e and atom combination strategies, with coherence threshold $\mu_{\max} = 0.7$.

atoms leading to $d(\Psi, \Phi) \gtrsim \frac{1}{2}$ is actually the best we can hope for.
Therefore, in the next section we will use our candidate atoms to make the big step towards adaptive selection of both sparsity level and dictionary size.

5. Adaptive dictionary learning

We first investigate how to adaptively choose the sparsity level for a dictionary of fixed size.

5.1 Adapting the sparsity level

In the numerical simulations of the last section we have seen that the sparsity level S given as parameter to the ITKrM algorithm influences both the convergence speed and the final precision of the learned dictionary.
When underestimating the sparsity level, meaning providing $S_e < S$ instead of S, the algorithm tends to recover the generating dictionary in less iterations than with the true sparsity level. Note also that 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, so that the estimate of every atom in I^t, even if not in 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 $a = Q_{I^t} r$ 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. Depending on the size of the noise and the inner product this might not always be beneficial to the final estimate. Indeed, we have seen that for large S, where the smallest coefficients in the support are already quite small, overestimating the support does not improve the final precision.
To further see that 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, clean signals. Then 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. 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. Indeed, 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, so 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

$$\mathbb{E}\left(\sharp\{k : |\langle r, \phi_k\rangle|^2 > \theta^2 \cdot \|r\|_2^2\}\right) = \sum_k \mathbb{P}\left(|\langle r, \phi_k\rangle|^2 > \theta^2 \cdot \|r\|_2^2\right) < 2(K-S)e^{-\frac{d\theta^2}{2}}. \tag{20}$$

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

$$\frac{1}{N}\sum_n \sharp\{k : |\langle a_n, \phi_k\rangle|^2 > \theta_K^2 \cdot \|a_n\|_2^2\} \lesssim \frac{1}{2}, \tag{21}$$

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

$$\frac{1}{N}\sum_n \sharp\{k : |\langle a_n, \phi_k\rangle|^2 > \theta_K^2 \cdot \|a_n\|_2^2\} \gtrsim \frac{m}{2}. \tag{22}$$

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 now 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,

$$\begin{aligned}\tilde{x}_{I^t} = \Phi_{I^t}^\dagger(\Phi_I x_I + r) &= \Phi_{I^t}^\dagger(\Phi_{I_{i\rightarrow}} x_{I_{i\rightarrow}} + \phi_i x_i + r)\\ &= (x_{I_{i\rightarrow}}, 0) + (\Phi_{I^t}^\star\Phi_{I^t})^{-1}\Phi_{I^t}^\star(\phi_i x_i + r),\end{aligned} \tag{23}$$

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. We know (from Appendix B or [7]) 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

$$\frac{1}{d}\|P(\Phi_{I^t})y\|_2^2 + \frac{2\log(4K)}{d}\|Q(\Phi_{I^t})y\|_2^2. \tag{24}$$

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

$$S_e^{new} = S_e + \operatorname{sign}(\bar{S} - S_e). \tag{25}$$

To avoid getting lost between numerical and explorative sections we will postpone an algorithmic summary to the appendix and testing of our adaptive sparsity selection to Subsection 5.3. Instead we next address the big question how to adaptively select the dictionary size.

5.2 Adapting the dictionary size

The common denominator of all popular dictionary learning algorithms, from MOD to K-SVD, is that before actually running them one has to choose a dictionary size. This choice 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 occurence 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 recovery precision, meaning that a bigger dictionary might not actually provide 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

$$\mathbb{P}\left(|\langle \phi_k, r\rangle| > \tau \|r\|_2\right) \leq 2\exp\left(-\frac{d\tau^2}{2}\right). \tag{26}$$

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 after replacing a double atom with the 1:1 complement $\phi_i - \phi_j$ of a 1:1 atom $\phi_i + \phi_j$, it takes 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 brings us 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

$$\mathbb{P}\left(|\langle \gamma_k, r\rangle| > \tau_\Gamma \|r\|_2\right) \leq 2\exp\left(-\frac{d\tau_\Gamma^2}{2}\right). \tag{27}$$

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, having dealt with all aspects necessary for making ITKrM adaptive, it is time to test whether adaptive dictionary learning actually works.

5.3 Experiments on synthetic data

We first test our adaptive dictionary learning algorithm on synthetic data[3]. The basic setup is the same as in Subsection 4.3. However, one type of training signals will again consist of 4, 6 and 8-sparse signals in a 1:2:1 ratio with 5% outliers, while the second type will consist of 8, 10 and 12-sparse signals in a 1:2:1 ratio with 5% outliers. Additionally, we will consider the following settings.
The **minimal number of reliable observations** M for a dictionary atom is set to either d, $\lfloor d \log d \rceil$ or $\lfloor 2d \log d \rceil$ with corresponding coefficient thresholds $\tau = \sqrt{2\log(2N/M)/d}$. For the candidate atoms the minimal number of reliable observations in the 4th (and last) candidate iteration is always set to $M_\Gamma = d$.

3. Again we want to point all interested in reproducing the experiments to the matlab toolbox available at `https://www.uibk.ac.at/mathematik/personal/schnass/code/adl.zip`

The **sparsity level** is adapted after every iteration starting with iteration $m = \lfloor \log d \rceil = 5$. The initial sparsity level is 1.
Promising candidate atoms are added to the dictionary after every iteration, starting again in the m-th iteration. In the last $3m$ iterations no more candidate atoms are added to the dictionary.
Coherent dictionary atoms are merged after every iteration, using the threshold $\mu_{\max} = 0.7$. As weights for the merging we use the value function of the atoms from the most recent iteration.
Unused dictionary atoms are pruned after every iteration starting with iteration $2m$. An atom is considered unused if in the last m iterations the number of reliable observations has always been smaller than M. Candidate atoms, freshly added to the dictionary, can only be deleted because they are unused at least m iterations later. In each iteration at most $\lfloor d/5 \rceil$ unused atoms are deleted, with an additional safeguard for very undercomplete dictionaries ($K_e < d/10$) that at most half of all atoms can be deleted.
The **initial dictionary** is chosen to be either of size $K_e = d = 128$, $K_e = 4d = 512$ or the correct size $K_e = K$, with the atoms drawn i.i.d. from the unit sphere as before. Figure 5 shows the results averaged over 10 trials each using a different initial dictionary.

The first observation is that all our effort paid off and that adaptive dictionary learning works. For the smaller average sparsity level $S = 6$, adaptive ITKrM always recovers all atoms of the dictionary and only overshoots and recovers more atoms for $M = d$. The main difference in recovery speed derives from the size of the initial dictionary, where a larger dictionary size leads to faster recovery.
For the more challenging signals with average sparsity level $S = 10$, the situation is more diverse. So while the initial dictionary size mainly influences recovery speed but less the final number of recovered atoms, the cut off threshold M for the minimal number of reliable observations is critical for full recovery. So for $M = d$ adaptive ITKrM never recovers the full dictionary. We can also see that not recovering the full dictionary is strongly correlated with overestimating the dictionary size. Indeed, the higher the overestimation factor for the dictionary size is, the lower is the amount of recovered atoms. For example, for $M = d$, $K_e = 512$ the dictionary size is overestimated by a factor 1.5 and only about half of the dictionary atoms are recovered. To see that the situation is not as bad as it seems we have a look at the average sorted atom recovery error. That is, we sort the recovery errors $(d(\phi_k, \Psi))_k$ after 100 iterations in ascending order and average over the number of trials. The resulting curves are depicted in Figure 6. As we can see, overestimating the dictionary size degrades the recovery in a gentle manner. For the unrecovered atoms in case $M = d$, $K_e = 512$, the largest inner product with an atom in the recovered dictionary is below the cut-off threshold of 0.99 which corresponds to an error of size ≈ 0.14 but for almost all of them it is still above 0.98 which corresponds to an error of 0.2. Also the oscillating recovery behaviour for $M = \lfloor d \log(d) \rceil$ before the final phase, where no more atoms are added, can be explained by the fact that the worst approximated atoms have average best inner product very close to 0.99. So, depending on the batch of training signals in each iteration, their inner product is below or above 0.99 and accordingly they count as recovered or not. In general, we can see that the more accurate the estimate of the dictionary size is, the better is the recovery precision of the learned dictionary. This is only to be expected. After all, whenever thresholding picks a superfluous atom instead of a correct atom, the number of

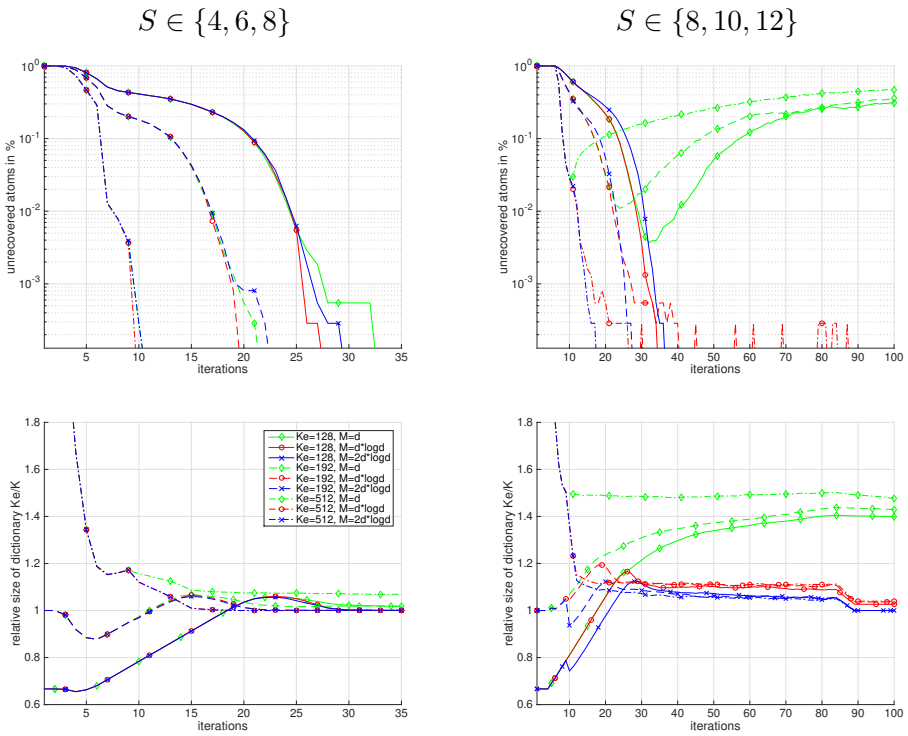

Figure 5: Average recovery rates (top row) and dictionary sizes (bottom row) for adaptive dictionary learning based on ITKrM on signals with sparsity $S = 4, 6, 8$ (left column) resp. $S = 8, 10, 12$ (right column) in a 1:2:1 ratio for various initial dictionary sizes K_e and required number of observations per atom M.

observations for the missing atom is reduced and moreover, the residual error added to the correctly identified atoms is increased.

The relative stability of these spurious atoms can in turn be explained by the fact that $S = 10$ is at the limit of admissible sparsity for a generating dictionary with $\mu(\Phi) = 0.32$, especially for sparse coefficients with a dynamic range of $0.9^{S-1} \approx 2.58$. In particular, thresholding is not powerful enough to recover the full support, so the residuals still contain several generating atoms. This promotes candidate atoms that are a sparse pooling of all dictionary atoms. These poolings again have a good chance to be selected in the thresholding and to be above the reliability threshold, thus positively reinforcing the effect. A quick look at the estimated sparsity level as well as the average number of coefficients above the threshold, or in other words, the average number of (probably) correctly identified atoms in the support, denoted by S_t, also supports this theory. So for average sparsity level $S = 6$ the estimated sparsity level is $S_e = 6 = \lfloor 5.7 \rceil$ and the average number of correctly identified atoms is $S_t \approx 5$, regardless of the setting. This is quite close to the average number of correctly identifiable atoms given $S_e = 6$, which is $0.95 * (0.25 * 4 + 0.75 * 6) = 5.225$. For average generating sparsity level $S = 10$ the table below lists $S_e : S_t$ for all settings. We can see that even for the settings where the full dictionary is recovered, the estimated

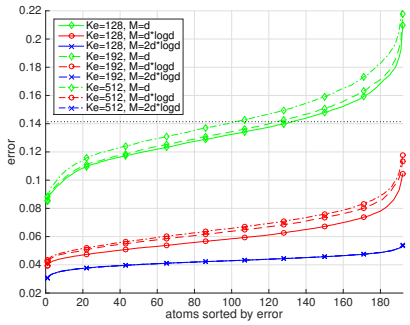

Figure 6: Average sorted recovery error $(d(\phi_k, \Psi))_k$ after 100 iterations of adaptive ITKrM on signals with sparsity $S = 8, 10, 12$ in a 1:2:1 ratio for various initial dictionary sizes K_e and required number of observations per atom M.

	d	$d\log(d)$	$2d\log(d)$
128	8 : 5.5	9 : 7.2	9 : 7.2
192	8 : 5.4	9 : 7.0	9 : 7.2
512	7 : 4.8	9 : 6.3	9 : 7.2

sparsity level is below 10 and the number of correctly identified atoms lags even more behind. For comparison, for $S_e = 9$ the average number of correctly identifiable atoms is $0.95 * (0.25 * 8 + 0.75 * 9) = 8.3125$.

We also want to mention that for signals with average generating sparsity $S = 6$ and $S_e = 6$ we can at best observe each atom $5.225 \cdot \frac{N}{K} \approx 3266$ times which is only about 5 times the threshold $d\log(d)$. For the signals with higher sparsity level and $S_e = 9$ we can at best observe each atom $8.3125 \cdot \frac{N}{K} \approx 5195$ times which is about 8 times the threshold $d\log(d)$, meaning that we are further from the critical limit where we would also remove an exactly recovered generating atom. In general, when choosing the minimal number of observations M, one needs to take into account that the number of recoverable atoms is limited by $K_{\max} \leq S_e * N/M$. On the other hand for larger S/S_e the generating coefficients will be smaller, meaning that they will be less likely to be over the threshold $\tau = \sqrt{2\log(2N/M)/d}$ if M is small. This suggests to also adapt M, τ in each iteration according to the current estimate of the sparsity level. Another strategy to reduce overshooting effects is to replace thresholding by a different approximation algorithm in the last rounds. The advantage of thresholding over more involved sparse approximation algorithms is its stability with respect to perturbations of the dictionary, see [32] for the case of OMP. The disadvantage is that it can only handle small dynamic coefficient ranges. However, we have seen that using thresholding, we can always get a reasonable estimate of the dictionary. Also in order to estimate S_e, we already have a good guess which atoms of the threshold support were correct and which atoms outside should have been included. This suggests to remove any atom from the support for which there is a more promising atom outside the support, or in other words, to update the support by thresholding $(\Psi_{I_t}^{\dagger} y, \Psi_{I_t^c}(\mathbb{I}_d - P(\Psi_I)y)$. Iterating this procedure until the support is stable is known as Hard Thresholding Pursuit (HTP),

[15]. Using 2 iterations of HTP would not overly increase the computational complexity of adaptive ITKrM but could help to weed out spurious atoms. Also by not keeping the S_e best atoms but only those above the threshold τ one could deal with varying sparsity levels, which would increase the final precision of the recovered dictionary.
Such a strategy might also help in addressing the only case where we have found our adaptive dictionary learning algorithm to fail spectacularly. This is - at first glance surprisingly - the most simple case of exactly 1-sparse signals and an initial dictionary size smaller than the generating size. At second glance it is not that surprising anymore. In case of underestimating the dictionary size $K - K_e = K_m > 0$ the best possible dictionary consists of $K - 2K_m$ generating atoms and K_m 1:1 combinations of 2 non-orthogonal atoms of the form $\phi_{ij} = (\phi_i + h\phi_j)/\alpha_{ij}$, where $h = \operatorname{sign}\langle\phi_i, \phi_j\rangle$ and $\alpha_{ij} = \sqrt{2 + 2\,|\langle\phi_i, \phi_j\rangle|}$. In such a situation the (non-zero) residuals are again 1-sparse in the 1:1 complements $\tilde{\phi}_{ij} = (\phi_i - h\phi_j)/\tilde{\alpha}_{ij}$, where $\tilde{\alpha}_{ij} = \sqrt{2 - 2\,|\langle\phi_i, \phi_j\rangle|}$, and so the replacement candidates will be the 1:1 complements. However, the problem is that $\tilde{\phi}_{ij}$ is never picked by thresholding since for both $y \approx \phi_i$ and $y \approx \phi_j$ the inner product with ϕ_{ij} is larger,

$$|\langle\phi_{ij}, \phi_i\rangle| = \sqrt{\frac{1 + |\langle\phi_i, \phi_j\rangle|}{2}} > \sqrt{\frac{1 - |\langle\phi_i, \phi_j\rangle|}{2}} = |\langle\tilde{\phi}_{ij}, \phi_i\rangle|. \tag{28}$$

Still the inner product of $\tilde{\phi}_{ij}$ with the residual has magnitude $\approx 1/2 > \tau$ and so would be included in the support in a second iteration of HTP, thus keeping the chance that the pair $(\phi_{ij}, \tilde{\phi}_{ij})$ rotates into the correct configuration (ϕ_i, ϕ_j) alive.
We will postpone a more in-depth discussion of how to further stabilise and improve adaptive dictionary learning to the discussion in Section 6. Here we will first check whether adaptive dictionary learning is robust to reality by testing it on image data.

5.4 Experiments on image data

In this subsection we will learn dictionaries for the images *Mandrill* and *Peppers*. For those interested in results on larger, more practically relevant datasets, we refer to [33], where our adaptive dictionary learning schemes are used for image reconstruction in accelerated 2D radial cine MRI.
The **training signals** are created as follows. Given a 256×256 image, we contaminate it with Gaussian noise of variance $\tilde{\rho}^2 = \rho^2/255$ for $\rho^2 \in \{0, 5, 10, 15, 20\}$. From the noisy image we extract all 8×8 patches (sub-images), vectorise them and remove their mean. In other words, we assume that the constant atom, $\phi_0 \equiv 1/8$, is always contained in the signal, remove its contribution and thus can only learn atoms that are orthogonal to it.
The **set-up** for adaptive dictionary learning is the same as for the synthetic data, taking into account that for the number of candidates L and the memory m, we have $L = m = \lfloor \log(d) \rceil = 4$, since the signals have dimension $d = 64$. Also based on the lesson learned on the more complicated data set with average sparsity level $S = 10$, we only consider as minimal number of observations $M = \lfloor d \log(d) \rceil$ and $M = 2d \log(d)$. The initial dictionary size K_e is either $8, 64$ or 256 and in each iteration we use all available signals, $N = 62001$. All results are averaged over 10 trials, each using a different initial dictionary and - where applicable - a different noise-pattern. In the first experiment we compare the sizes of the dictionaries learned on both clean images with various parameter settings as well as their

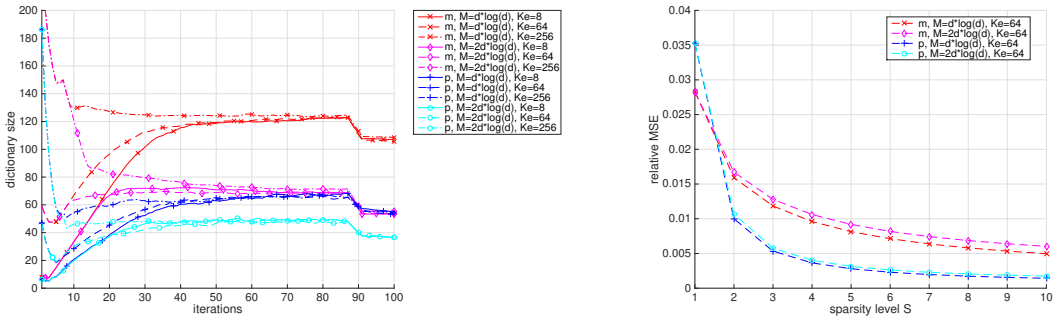

Figure 7: Average dictionary sizes for adaptive dictionary learning based on ITKrM on all patches of *Mandrill*/*Peppers* for various initial dictionary sizes K_e and required number of observations per atom M (left). Average sparse approximation error of all patches of *Mandrill*/*Peppers* using OMP and the learned dictionaries with $K_e = 64$ and both choices of M (right).

approximation powers. The approximation power of a dictionary augmented by the flat atom ϕ_0 for a given sparsity level S is measured as $\|Y - \tilde{Y}\|_F^2 / \|Y\|_F^2$, where $\tilde{Y} = (\tilde{y}_1, \dots, \tilde{y}_n)$ and $\tilde{y}_n$ is the S-sparse approximation to y_n calculated by Orthogonal Matching Pursuit, [34]. The results are shown in Figure 7. We can see that as for synthetic data the final size of the learned dictionary does not depend much on the initial dictionary size, but does depend on the minimal number of observations. So for $M = \lfloor d \log(d) \rceil$ the average dictionary size is about 106 atoms for *Mandrill* and 55 atoms for *Peppers*, while for $M = 2d \log(d)$ we have about 54 atoms on *Mandrill* and 36 atoms on *Peppers*. The estimated sparsity level vs. average number of correctly identified atoms for *Mandrill* is $S_e = \lfloor 2.1 \rceil = 2$ vs. $S_t \approx 1.5$ and for *Peppers* $S_e = \lfloor 2.9 \rceil = 3$ vs. $S_t \approx 2.25$. Comparing the approximation power, we see that for both images the smaller (undercomplete) dictionaries barely lag behind the larger dictionaries. The probably most interesting aspect is that despite being smaller, the *Peppers*-dictionaries lead to smaller error than the *Mandrill*-dictionaries. This confirms the intuition that the smooth image *Peppers* has a lot more sparse structure than the textured image *Mandrill*. To better understand why for both images the larger dictionaries do not improve the approximation much, we have a look at the number of reliable observations for each atom in the last trial of $K_e = 64$ in Figure 8. The corresponding dictionaries for *Mandrill*/*Peppers* can be found in Figures 9/10.

We can see that for both images the number of times each atom is observed strongly varies. If we interpret the relative number of observations as probability of an atom to be used, the first ten atoms are more than 10 times more likely to be used/observed than the last ten atoms. This accounts for the fact that by increasing the threshold for the minimal number of observations, we reduce the dictionary size without much affecting the approximation power.

In our second experiment we learn adaptive dictionaries on *Mandrill* contaminated with Gaussian noise of variance $\tilde{\rho}^2 = \rho^2/255$ for $\sigma^2 \in \{5, 10, 15, 20\}$, corresponding to average peak signal to noise ratios $\{34.15, 28.13, 24.61, 22.11\}$. We again compare their sizes and

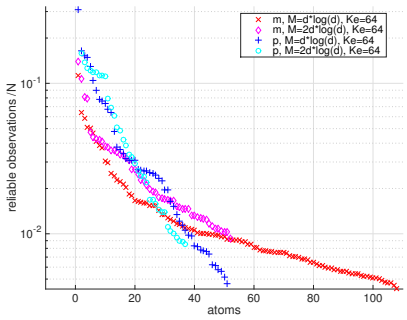

Figure 8: Final number of reliable observations of the atoms in the dictionaries learned on *Mandrill/Peppers* with initial dictionary size $K_e = 64$ in the last trial.

their approximation power for the clean image patches. The results are shown in Figure 12 and two example dictionaries are shown in Figure 11.

We can see that the size of the dictionary decreases quite drastically with increasing noise, while the approximation power degrades only very gently. The sparsity level chosen by the algorithm is $S_e = \lfloor 1.80 \rceil = 2$ for $\sigma_2 = 5$ and $S_e = \lfloor 1.11 \rceil = \lfloor 0.89 \rceil = 1$ for $\sigma_2 \in \{15, 20\}$. For $\sigma^2 = 10$ the average recoverable sparsity level is ≈ 1.5 so that for all trials the estimated sparsity level S_e alternates between 1 and 2 in consecutive iterations. The fact that with increasing noise both the dictionary size and the sparsity level decrease but not the approximation power indicates that less often used atoms also tend to capture less energy per observation. This suggests an alternative value function, where each reliable observation is additionally weighted, for instance, by the squared coefficient or inner product. The corresponding cut-off threshold then is the minimal amount of energy that a reliable atom needs to capture from the training signals. Such an alternative value function could be useful in applications like dictionary based denoising, where every additional atom not only leads to better approximation of the signals but also of the noise.
Now that we have seen that adaptive dictionary learning produces sensible and noise robust results not only on synthetic but also on image data, we will turn to a final discussion of our results.

6. Discussion

In this paper we have studied the global behaviour of the ITKrM (Iterative Thresholding and K residual means) algorithm for dictionary learning. We have proved that ITKrM contracts a dictionary estimate Ψ towards the generating dictionary Φ whenever the cross-Gram matrix $\Psi^\star\Phi$ is diagonally dominant and Ψ is incoherent and well-conditioned. Further, we have identified dictionaries, that are not equivalent to a generating dictionary but seem to be stable fixed points of ITKrM. Using our insights that these fixed points always contain several atoms twice, meaning they are coherent, and that the residuals contain information about the missing atoms, we have developed a heuristic for finding good candidates which we can use to replace one of two coherent atoms in a dictionary estimate. Simulations on

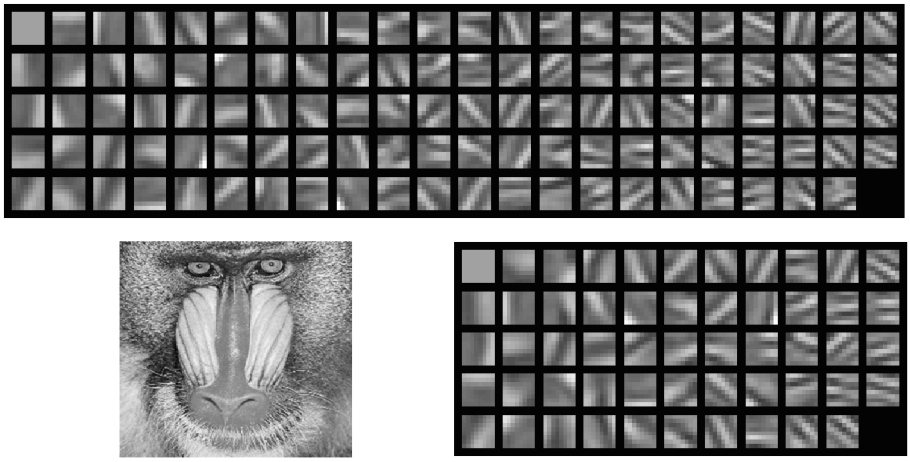

Figure 9: Dictionaries learned on *Mandrill* with initial dictionary size $K_e = 64$ and required number of observations $M = \lfloor d \log(d) \rceil$ (top) resp. $M = 2d \log(d)$ (bottom).

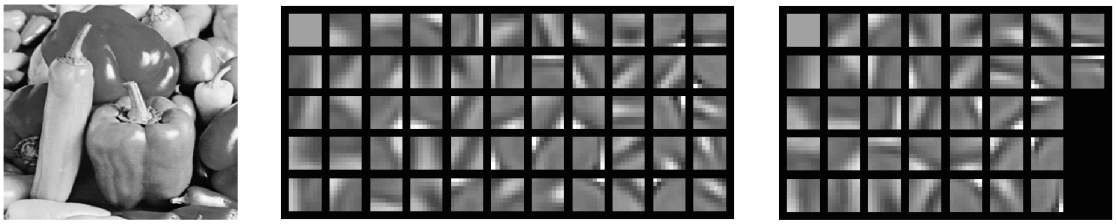

Figure 10: Dictionaries learned on *Peppers* with initial dictionary size $K_e = 64$ and required number of observations $M = \lfloor d \log(d) \rceil$ (middle) resp. $M = 2d \log(d)$ (right).

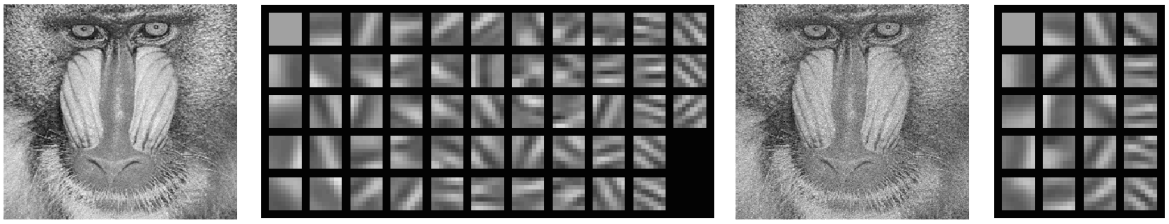

Figure 11: Dictionaries learned on the *Mandrill* image contaminated with Gaussian noise of variance $\tilde{\rho}^2 = 10/255$ (left) and $\tilde{\rho}^2 = 20/255$ (right), initial dictionary size $K_e = 64$ and required number of observations $M = \lfloor d \log(d) \rceil$.

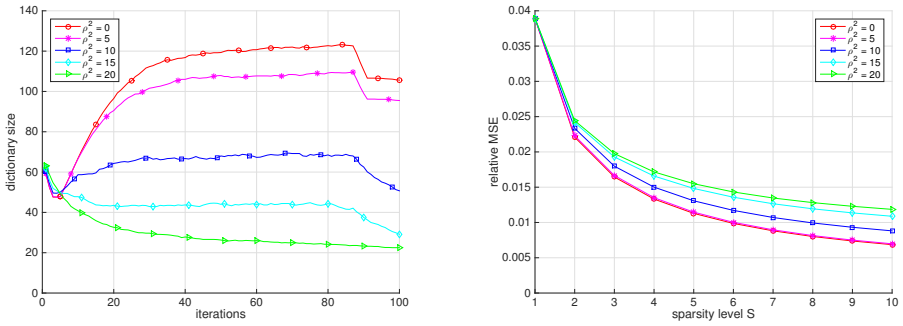

Figure 12: Average dictionary sizes for adaptive dictionary learning based on ITKrM on all patches of the *Mandrill* image contaminated with Gaussian noise of variance $\tilde{\rho}^2 = \rho^2/255$, initial dictionary size $K_e = 64$ and required number of observations $M = \lfloor d \log(d) \rceil$ (left). Corresponding average sparse approximation error of all clean patches of *Mandrill* using OMP and the learned dictionaries. (right).

synthetic data have shown that replacement using these candidates improved over random or no replacement, always leading to recovery of the full dictionary.

Armed with replacement candidates, we have addressed one of the most challening problems in dictionary learning - how to automatically choose the sparsity level and dictionary size. We have developed a strategy for adapting the sparsity level from the initial guess $S_e = 1$ and the dictionary size by decoupling replacement into pruning of coherent and unused atoms and adding of promising candidates. The resulting adaptive dictionary learning algorithm has been shown to perform very well in recovering a generating dictionary from random initialisations with various sizes on synthetic data with sparsity levels $S \geq 2$ and in learning meaningful dictionaries on image data.

Note that our strategy for learning replacement candidates and adaptivity can be easily adapted to any other alternating minimisation algorithm for dictionary learning, such as MOD or K-SVD, [13, 3]. Instead of using ITKsM on the residuals one simply has to use a small scale version with sparsity level $S = 1$ of the respective algorithm. Preliminary experiments on synthetic data show that in the case of K-SVD and MOD this gives very similar results, [39]. Interestingly, however, it is much more difficult to stabilise the adaptive version of K-SVD on image data. The reason for this seems to be that the more sophisticated sparse approximation routine Orthogonal Matching Pursuit (OMP), [34], promotes the convergence of candidate atoms to high energy subspaces. This leads to a concentration of many coherent atoms in these subspaces, without proportional increase in the approximation quality of the dictionary. These experimental findings are supported by theory, showing that while an excellent choice if the sparsifying dictionary is known, the performance of OMP rapidly degrades if the sparsifying dictionary is perturbed, [32]. Indeed on synthetic data without replacement procedures, K-SVD with OMP produces more double atoms than ITKrM. However, when using thresholding as sparse approximation procedures also for K-SVD and MOD, MOD produces the smallest number of double atoms followed by K-SVD and then ITKrM. With replacement MOD recovers the most accurate dictionary

again followed by K-SVD and ITKrM.
These observations suggest that the use of more sophisticated algorithms to bridge the gap between average sparsity level $\bar{S}$ and the average number of correctly identified atoms S_t is only advisable in the last iterations when the number of atoms is not changed anymore.
A promising direction to avoid the concentration of coherent atoms in high energy subspaces is to use a different value function for the dictionary atoms. The one proposed here is based on the assumption that all atoms are used equally often, which for image data was clearly not the case. Alternatively, the current function could be scaled, that is, every reliable occurence of an atom is weighted with the squared coefficient which is related to the signal energy lost without this atom. Preliminary experiments with ITKrM using this weighted value function and an accordingly scaled cut-off indicate that it indeed helps to remove spurious atoms on synthetic data and leads to smaller dictionaries with the same approximation power on image data.
Another interesting question, which leads directly to our future theoretical research directions, is how to combine the value function of an atom with its coherence. The key to a solution lies in the analysis of dictionary learning from data where not all atoms are used equally often. A first step in this direction was laid in [37], which provides results for the conditioning of random subdictionaries with non-homogeneous atom distributions together with an analysis for several sparse approximation algorithms. It also revealed the interplay between coherence structure of the dictionary and the probability of each atom to be used. Also to better model such a situation it will become necessary to look at a modification of thresholding, which uses as similarly noise inspired threshold τ as our value function, rather than a fixed sparsity level.
To put our replacement strategies on a more solid foundation, we are currently working on proofs showing the existence of stable spurious fixed points of ITKrM. The next steps are to show that all contractive areas are actually convergent as well as partial convergence of most atoms as long as the cross-Gram-matrix has the desired structure except for a few rows or columns. Finally, we want to transfer our characterisation of contractive areas to MOD and weighted ITKrM, where we replace the sign of the inner products in the update formula with the inner products themselves. This weighted version can also be seen as approximative K-SVD in the sense that only one power iteration is used to calculate the left singular vectors.

Acknowledgments

This work was supported by the Austrian Science Fund (FWF) under Grant no. Y760. The computational results presented have been achieved (in part) using the HPC infrastructure LEO of the University of Innsbruck. We would like to thank Alexander Steinicke for his explanations concerning Freedman's inequality, quadratic variation and the Doob martingale as well as Simon Ruetz for proof-reading the manuscript.

Appendix A. Exact Statement and Proof of Theorem 1

We only state and prove the exact version of the second part, since the first part consists literally of considering just one iteration of ITKrM and replacing in Theorem 4.2 of [43] the assumption on the distance $d(\Psi, \Phi)$ with the assumptions on the coherence and the operator norm of Ψ, ie.

$$d(\Psi,\Phi) \leq \frac{1}{32\sqrt{S}} \quad \rightsquigarrow \quad \mu(\Psi) \leq \frac{1}{20\log K} \quad \text{and} \quad \|\Psi\|_{2,2}^2 \leq \frac{K}{134e^2 S\log K} - 1. \tag{29}$$

Similarly the amendment to the proof consists in using Lemma 8 in Appendix B.2 instead of Lemma B.8 of [43] and potentially some tweaking of constants.

Proposition 2 (Theorem 1(b) exact) *Assume that the signals y_n follow model (7) 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,*

$$\mu(\Psi) \leq \frac{1}{20\log K} \quad \textit{and} \quad \|\Psi\|_{2,2}^2 \leq \frac{K}{134e^2 S\log K} - 1. \tag{30}$$

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

$$\begin{aligned}\min_k |\langle \psi_k, \phi_k\rangle| \geq \max\Big\{ & 8\,\gamma_{gap}\cdot \max_k |\langle \psi_k,\phi_k\rangle|\,, \\ & 40\,\gamma_\rho \cdot \sqrt{\log K}, \\ & 48\,\gamma_{dyn}\cdot \log K\cdot \mu(\Phi,\Psi), \\ & 14\,\gamma_{dyn}\cdot \sqrt{\|\Phi\|_{2,2}^2 S\log K/(K-S)}\Big\},\end{aligned} \tag{31}$$

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

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

Proof We follow the outline of the proof for Theorem 4.2 in [43]. However, to extend the convergence radius we need to introduce new ideas, first for bounding the difference between the oracle residuals based on Ψ and Φ, replacing Lemma B.8 of [43], and second for bounding the probability of thresholding with Ψ not recovering the generating support or preserving the generating sign, replacing Lemma B.3/4 of [43]. We denote the thresholding residual based on Ψ by

$$R^t(\Psi, y_n, k) := \big[y_n - P(\Psi_{I^t_{\Psi,n}})y_n + P(\psi_k)y_n\big]\cdot \operatorname{sign}(\langle\psi_k, y_n\rangle)\cdot\chi(I^t_{\Psi,n},k), \tag{32}$$

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

$$R^o(\Psi, y_n, k) := \big[y_n - P(\Psi_{I_n})y_n + P(\psi_k)y_n\big] \cdot \sigma_n(k) \cdot \chi(I_n, k). \tag{33}$$

Abbreviating $s_k = \frac{1}{N}\sum_n \langle y_n, \phi_k\rangle \cdot \sigma_n(k) \cdot \chi(I_n, k)$ and setting $B := \|\Phi\|_{2,2}^2$ as well as $\varepsilon := d(\Psi, \Phi)$ for conciseness, we know from the proof of Theorem 4.2 in [43] that

$$\begin{aligned}\|\bar{\psi}_k - s_k\phi_k\|_2 \leq \frac{1}{N}\Big\| \sum_n \big[R^t(\Psi, y_n, k) - R^o(\Psi, y_n, k)\big] \Big\|_2 \\ + \frac{1}{N}\Big\| \sum_n \big[R^o(\Psi, y_n, k) - R^o(\Phi, y_n, k)\big] \Big\|_2 \\ + \frac{1}{N}\Big\| \sum_n \big[y_n - P(\Phi_{I_n})y_n\big] \cdot \sigma_n(k) \cdot \chi(I_n, k)\Big\|_2. \end{aligned} \tag{34}$$

By Lemma B.6 from [43] we have

$$\begin{aligned}\mathbb{P}\left(\left|\frac{1}{N}\sum_n \chi(I_n, k)\sigma_n(k)\langle y_n, \phi_k\rangle\right| \leq (1-t_0)\frac{C_r\gamma_{1,S}}{K}\right) \\ \leq \exp\left(-\frac{NC_r^2\gamma_{1,S}^2 \cdot t_0^2}{2K(1 + \frac{SB}{K} + S\rho^2 + t_0 C_r\gamma_{1,S}\sqrt{B+1}/3)}\right). \end{aligned} \tag{35}$$

By Lemma 6 in Appendix B.1 (substituting Lemma B.3/4 of [43]) we have

$$\begin{aligned}\mathbb{P}\left(\frac{1}{N}\left\|\sum_n \big[R^t(\Psi, y_n, k) - R^o(\Psi, y_n, k)\big]\right\|_2 > \frac{18(S+1)\sqrt{B+1}}{K^3} + \frac{C_r\gamma_{1,S}}{K}t_1\varepsilon\right) \\ \leq 2\exp\left(-\frac{NC_r^2\gamma_{1,S}^2 t_1^2\varepsilon^2}{\frac{108(S+1)(B+1)}{K} + 3t_1\varepsilon C_r\gamma_{1,S}K\sqrt{B+1}}\right). \end{aligned} \tag{36}$$

By Lemma 8 in Appendix B.2 (substituting Lemma B.8 of [43]) we have that for $0 \leq t_2 \leq 1/8$

$$\begin{aligned}\mathbb{P}\left(\frac{1}{N}\left\|\sum_n [R^o(\Psi, y_n, k) - R^o(\Phi, y_n, k)]\right\|_2 \geq \frac{C_r\gamma_{1,S}}{K}(0.308\varepsilon + t_2\varepsilon)\right) \\ \leq \exp\left(-\frac{NC_r^2\gamma_{1,S}^2 \cdot t_2^2\varepsilon}{12K\max\{S, B\}^{\frac{3}{2}}} + \frac{1}{4}\right), \end{aligned} \tag{37}$$

and by Lemma B.7 from [43] we have

$$\begin{aligned}\mathbb{P}\left(\left\|\frac{1}{N}\sum_n \big[y_n - P(\Phi_{I_n})y_n\big] \cdot \sigma_n(k) \cdot \chi(I_n, k)\right\|_2 \geq \frac{C_r\gamma_{1,S}}{K}\, t_3\varepsilon\right) \\ \leq \exp\left(-\frac{NC_r^2\gamma_{1,S}^2 \cdot t_3\varepsilon}{8K\max\{S, B+1\}}\min\left\{\frac{t_3\varepsilon}{(1-\gamma_{2,S}+d\rho^2)}, 1\right\} + \frac{1}{4}\right). \end{aligned} \tag{38}$$

Thus, with high probability we have $s_k \geq (1-t_0)\frac{C_r\gamma_{1,S}}{K}$ and

$$\left\|\bar{\psi}_k - s_k\phi_k\right\|_2 \leq \frac{C_r\gamma_{1,S}}{K}\left(\frac{18(S+1)\sqrt{B+1}}{K^2 C_r\gamma_{1,S}\varepsilon} + t_1 + 0.308 + t_2 + t_3\right)\varepsilon. \tag{39}$$

Note that we only need to take into account distances $\varepsilon > \frac{1}{32\sqrt{S}}$, 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 $\gamma_{1,S} \geq Sc(S)$. Combining this with the bound on C_r in (10) we get

$$\frac{1}{C_r\gamma_{1,S}} \leq \frac{\sqrt{1+5d\rho^2}}{(1-e^{-d})\gamma_{1,S}} \leq \frac{\sqrt{1+3\log K}}{(1-e^{-d})} + \frac{\rho}{c(S)}\frac{\sqrt{5d}}{S(1-e^{-d})}. \tag{40}$$

The conditions in (31) 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

$$\frac{18(S+1)\sqrt{B+1}}{K^2 C_r\gamma_{1,S}\varepsilon} \leq 0.025, \tag{41}$$

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

$$\max_k \left\|\bar{\psi}_k - s_k\phi_k\right\|_2 \leq 0.633 \cdot \frac{C_r\gamma_{1,S}}{K}\varepsilon \quad \text{and} \quad \min_k s_k \geq 0.95 \cdot \frac{C_r\gamma_{1,S}}{K}, \tag{42}$$

which by Lemma B.10 from [43] implies that

$$d(\bar{\Psi},\Phi)^2 = \max_k \left\|\frac{\bar{\psi}_k}{\|\bar{\psi}_k\|_2} - \phi_k\right\|_2^2 \leq 2\left(1-\sqrt{1-\frac{0.633^2\varepsilon^2}{0.95^2}}\right) \leq \frac{2\cdot 0.633^2\varepsilon^2}{0.95^2} \leq 0.89\varepsilon^2, \tag{43}$$

except with probability

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

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

Appendix B. Technical Lemmata

Here we state the proofs of the two lemmata characterising the difference between the thresholding and the oracle residual resp. the difference between the oracle residuals based on the generating dictionary and a perturbation.

B.1 Difference between thresholding and oracle residual

To prove Lemma 6 we will make use of the scalar version of Bernstein's inequality [7] and Hoeffding's inequality [20]. We will also need Proposition 5, which is based on a version of Freedman's inequality, [16], to deal with sums of dependent random variables.

Theorem 3 (Scalar Bernstein, [7]) *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*

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

To prove Proposition 5 we need the following simplified version of Freedman's inequality.

Theorem 4 (Freedman, [16]) *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 5 *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 $t \geq 0$,*

$$\mathbb{P}\left(|\sum_{k=1}^{S} c_k \varepsilon_k v_{i_k}| \geq t\right) \leq 2\exp\left(\frac{-t^2}{2(\|c\|_\infty \|v\|_\infty t + \|c\|_2^2 \cdot \|v\|_2^2/(K - S))}\right). \tag{44}$$

Proof We will use Theorem 4 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

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

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. ∎

Now we are ready to prove the lemma estimating the error originating from thresholding failing to recover the generating supports and signs.

Lemma 6 *Assume that the signals y_n follow model (7) 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*

$$\begin{aligned}\min_k |\langle \psi_k, \phi_k \rangle| \geq \max \Big\{ &8\, \gamma_{gap} \cdot \max_k |\langle \psi_k, \phi_k \rangle|, \\ &40\, \gamma_\rho \cdot \sqrt{\log K}, \\ &48\, \gamma_{dyn} \cdot \log K \cdot \mu(\Phi, \Psi), \\ &14\, \gamma_{dyn} \cdot \sqrt{\|\Phi\|_{2,2}^2 S \log K/(K-S)} \Big\},\end{aligned} \tag{45}$$

then

$$\begin{aligned}\mathbb{P}\left(\frac{1}{N}\left\|\sum_n \left[R^t(\Psi, y_n, k) - R^o(\Psi, y_n, k)\right]\right\|_2 > \frac{18(S+1)\sqrt{\|\Phi\|_{2,2}^2+1}}{K^3} + \frac{C_r \gamma_{1,S}}{K} t\varepsilon\right) \\ \leq 2 \exp\left(-\frac{N C_r^2 \gamma_{1,S}^2 t^2 \varepsilon^2}{\frac{108(S+1)(\|\Phi\|_{2,2}^2+1)}{K} + 3t\varepsilon C_r \gamma_{1,S} K \sqrt{\|\Phi\|_{2,2}^2+1}}\right).\end{aligned} \tag{46}$$

Proof Throughout the proof we will 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

$$\Delta = \Big\| \sum_n \big[R^t(\Psi, y_n, k) - R^o(\Psi, y_n, k)\big] \Big\|_2 = \Big\| \sum_{n \in \mathcal{F} \cup \mathcal{G}} \big[R^t(\Psi, y_n, k) - R^o(\Psi, y_n, k)\big] \Big\|_2. \quad (47)$$

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

$$\Delta \leq \sum_{n \in \mathcal{F} \cup \mathcal{G}} \big(\|R^t(\Psi, y_n, k)\|_2 + \|R^o(\Psi, y_n, k)\|\big) \leq (2|\mathcal{F}| + |\mathcal{G}|)\sqrt{B+1}. \quad (48)$$

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

$$F = \left\{y : k \in I \wedge \big(I^t \neq I \vee \operatorname{sign}(\langle \psi_k, y \rangle) \neq \sigma(k)\big)\right\}, \quad (49)$$

which leads to

$$\mathbb{P}(|\mathcal{F}| \geq N\mathbb{P}(F) + Nt) \leq \exp\left(-\frac{t^2 N}{2\mathbb{P}(F) + t}\right). \quad (50)$$

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

$$\mathbb{P}(|\mathcal{G}| \geq N\mathbb{P}(G) + Nt) \leq \exp\left(-\frac{t^2 N}{2\mathbb{P}(G) + t}\right). \quad (51)$$

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 will show that with high probability for a signal y following the model in (7) 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 $I^t = I$, this means the recovery of all $i \in I$, we need to have

$$\min_{i \in I} |\langle \psi_i, y \rangle| > \max_{i \notin I} |\langle \psi_i, y \rangle|. \quad (52)$$

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 (remember that $\alpha_{\min} \leq |\langle \psi_i, \phi_i\rangle| \leq \alpha_{\max}$),

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

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

$$\big|\sum_{j\neq i}\sigma(j)c(p(j))\langle \psi_i, \phi_j\rangle\big| < \theta_1 \cdot c(S)\alpha_{\min} \quad \text{and} \quad |\langle \psi_i, r\rangle| < \theta_2 \cdot c(S)\alpha_{\min}, \tag{53}$$

where θ_1 and θ_2 ensure that

$$c(S)\alpha_{\min} - \theta_1 c(S)\alpha_{\min} - \theta_2 c(S)\alpha_{\min} \overset{!}{\geq} c(S+1)\alpha_{\max} + \theta_1 c(S)\alpha_{\min} + \theta_2 c(S)\alpha_{\min}. \tag{54}$$

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

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

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

$$\begin{aligned}
&\mathbb{P}\left(\left[I^t \neq I \vee \operatorname{sign}(\langle \psi_k, y\rangle) \neq \sigma(k)\right] \wedge k \in I\right) \\
&\qquad \leq \frac{1}{K}\sum_i \sum_{\ell \in \mathbb{S}} \mathbb{P}\Big(\Big|\sum_{j\neq i} \sigma(j)c(p(j))\langle \psi_i, \phi_j\rangle\Big| \geq \theta_1 c(S)\alpha_{\min}\Big| p(k) = \ell\Big) \\
&\qquad\qquad + \frac{S}{K}\sum_i \mathbb{P}\big(|\langle \psi_i, r\rangle| \geq \theta_2 c(S)\alpha_{\min}\big) \\
&\qquad \leq \frac{1}{K}\sum_i \sum_{\ell \in \mathbb{S}} \mathbb{P}\Big(\Big|\sum_{j\neq i} \sigma(j)c(p(j))\langle \psi_i, \phi_j\rangle\Big| \geq \theta_1 c(S)\alpha_{\min}\Big| p(k) = \ell\Big) \\
&\qquad\qquad + 2S\exp\left(\frac{-(\theta_2 c(S)\alpha_{\min})^2}{2\rho^2}\right). \qquad (55)
\end{aligned}$$

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 $j \in I\setminus\{i,k\}$, that captures most of the energy, and the other over $j \in (I^c \cup \{k\})\setminus\{i\}$. For $m_1 \in (0,1)$ and $m_2 = 1-m_1$ we have

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

The first term we estimate using Proposition 5 and the second term using Hoeffding's inequality. With some small simplifications we get for all i, including k,

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

Substituting the expression above into (55) we get

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

where θ_1 and θ_2 have to ensure (54) and $m_1 + m_2 = 1$. From this, whenever

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

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

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

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 will 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

$$\min_{i \in I} |\langle\psi_i, y\rangle| > |\langle\psi_k, y\rangle|. \tag{56}$$

Using the bounds from step 1,

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

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

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

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

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

To have this probability sufficiently small, we need to have

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

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

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

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

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

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 (50) and (51), we get

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

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

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

■

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.

B.2 Difference between oracle residuals

For the proof of Lemma 8 we will use the vector version of Bernstein's inequality.

Theorem 7 (Vector Bernstein, [24, 19, 25]) *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*

$$\mathbb{P}\left(\left\|\sum_n v_n - \sum_n \mathbb{E}(v_n)\right\|_2 \geq t\right) \leq \exp\left(-\frac{t^2}{8m_2} + \frac{1}{4}\right), \tag{57}$$

and, in general,

$$\mathbb{P}\left(\left\|\sum_n v_n - \sum_n \mathbb{E}(v_n)\right\|_2 \geq t\right) \leq \exp\left(-\frac{t}{8}\cdot\min\left\{\frac{t}{m_2}, \frac{1}{M+m_1}\right\} + \frac{1}{4}\right). \tag{58}$$

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 $m_2 = t(M+m_1)$.

We next 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 8 *Assume that the signals y_n follow the random model in (7). 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*

$$\mu(\Psi) \leq \frac{1}{20\log K} \quad \textit{and} \quad \|\Psi\|_{2,2}^2 \leq \frac{K}{134e^2 S\log K} - 1. \tag{59}$$

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

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

Proof Throughout the proof we will use the abbreviations $B = \|\Phi\|_{2,2}^2$ and $\bar{B} = \|\Psi\|_{2,2}^2$. We apply Theorem 7 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 [43] we know 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 that

$$\mathbb{E}(v) = \frac{C_r\gamma_{1,S}}{K}\binom{K-1}{S-1}^{-1}\sum_{|I|=S,k\in I}\big[P(\psi_k) - P(\Psi_I)\big]\phi_k. \tag{60}$$

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

$$\mathbb{E}(v) = \frac{C_r\gamma_{1,S}}{K}\binom{K-1}{S-1}^{-1}\sum_{|I|=S,k\in I} -P(\Psi_I)Q(\psi_k)\phi_k. \tag{61}$$

Since the perturbed dictionary Ψ is well-conditioned and incoherent, for most I the sub-dictionary Ψ_I will be a quasi isometry and $P(\Psi_I) \approx \Psi_I\Psi_I^\star$. 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

$$\|\mathbb{E}(v)\|_2 \leq \frac{C_r \gamma_{1,S}}{K} \left[\delta_0 + \mathbb{P}\big(\delta(\Psi_I) > \delta_0 \big| |I| = S, k \in I\big) \cdot (\bar{B}+1) + \frac{(S-1)\bar{B}}{K-1} \right] \varepsilon. \tag{62}$$

To estimate the probability of a subdictionary being ill-conditioned we use Chretien and Darses's results on the conditioning of random subdictionaries, which are slightly cleaner and thus easier to handle than the original results by Tropp, [50]. Theorem 3.1 of [10] reformulated for our purposes and applied to Ψ states that

$$\mathbb{P}\big(\delta(\Psi_I) > \delta_0 \big| |I| = S\big) \leq 216K \exp\left(- \min\left\{ \frac{\delta_0}{2\mu(\Psi)}, \frac{\delta_0^2 K}{4e^2 S\bar{B}} \right\} \right). \tag{63}$$

Together with the union bound,

$$\mathbb{P}\big(\delta(\Psi_I) > \delta_0 \big| |I| = S, k \in I\big) \leq \frac{K}{S} \cdot \mathbb{P}\big(\delta(\Psi_I) > \delta_0 \big| |I| = S\big), \tag{64}$$

this leads to

$$\|\mathbb{E}(v)\|_2 \leq \frac{C_r \gamma_{1,S}}{K} \left[\delta_0 + \frac{216K^2(\bar{B}+1)}{S} \exp\left(- \min\left\{ \frac{\delta_0}{2\mu(\Psi)}, \frac{\delta_0^2 K}{4e^2 S\bar{B}} \right\} \right) + \frac{S\bar{B}}{K} \right] \varepsilon. \tag{65}$$

Choosing $\delta_0 = 3/10$, 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

$$\|\mathbb{E}(v)\|_2 \leq 0.308 \cdot \frac{C_r \gamma_{1,S}}{K} \cdot \varepsilon, \tag{66}$$

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 [43] we get that

$$\mathbb{E}(\|v\|_2^2) \leq \mathbb{E}_p \left(\chi(I,k) \left[4\gamma_{2,S}\varepsilon^2 + \left(\frac{B(1-\gamma_{2,S})}{K-S} + \rho^2 \right) \|T(I,k)\|_F^2 \right] \right). \tag{67}$$

Since we are only interested in the regime $\varepsilon > O(1/\sqrt{S})$ we will accept an additional factor S in the final sample complexity in return for a crude but painless estimate. Concretely, we use that $T(I,k)$ is the difference of two orthogonal projections onto subspaces of dimension $S-1$, namely $P(\Phi_I) - P(\phi_k)$ and $P(\Psi_I) - P(\psi_k)$. This leads to the bound $\|T(I,k)\|_F^2 \leq 2(S-1) \leq 2S$ and we get

$$\mathbb{E}(\|v\|_2^2) \leq \frac{S}{K} \left(4\gamma_{2,S}\varepsilon^2 + \frac{2BS}{K-S}(1-\gamma_{2,S}) + 2S\rho^2 \right) \leq \frac{S}{K}(4\varepsilon^2 + 1/24), \tag{68}$$

where for the second 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 $\|v\|_2 \leq 2\sqrt{B+1}$,

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

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{N}\left\|\sum_n [R^o(\Psi, y_n, k) - R^o(\Phi, y_n, k)]\right\|_2 \geq \frac{C_r\gamma_{1,S}}{K}(0.308\varepsilon + t\varepsilon)\right) \\
&\leq \exp\left(-\frac{NC_r\gamma_{1,S}t\varepsilon}{8K}\min\left\{\frac{C_r\gamma_{1,S}t\varepsilon}{S(4\varepsilon^2+1/24)}, \frac{1}{\varepsilon+2\sqrt{B+1}}\right\} + \frac{1}{4}\right) \\
&\leq \exp\left(-\frac{NC_r^2\gamma_{1,S}^2t^2\varepsilon}{8K}\min\left\{\frac{1}{S(4\varepsilon+(24\varepsilon)^{-1})}, \frac{1}{3t\gamma_{1,S}\sqrt{B+1}}\right\} + \frac{1}{4}\right) \\
&\leq \exp\left(-\frac{NC_r^2\gamma_{1,S}^2t^2\varepsilon}{8K\max\{S,B\}}\min\left\{\frac{1}{4\varepsilon+(24\varepsilon)^{-1}}, \frac{1}{3t\sqrt{2}}\right\} + \frac{1}{4}\right) \\
&\leq \exp\left(-\frac{NC_r^2\gamma_{1,S}^2t^2\varepsilon}{12K\max\{S,B\}^{\frac{3}{2}}} + \frac{1}{4}\right).
\end{aligned}$$

■

Appendix C. Pseudocode

Algorithm C.1: ITKrM augmented for **r**eplacement/**a**daptivity - one iteration

a/r **Input:** Ψ, Y, S, Γ, M ; // dictionary, signals, sparsity, candidates, // minimal observations (only for adaptive)

Set: $m = \lfloor \log d \rceil$, $N_\Gamma = \lfloor N/m \rfloor$;
Initialise: $\bar{\Psi} = 0$, $\bar{\Gamma} = 0$, $\bar{S} = 0$;

foreach n **do**

// basic ITKrM steps
$I_n^t = \arg\max_{I:|I|=S} \|\Psi_I^\star y_n\|_1$; // thresholding
$x_n = \Psi_{I_n^t}^\dagger y_n$; // sparse coefficients
$a_n = y_n - \Psi_{I_n^t} x_n$; // residual

r $\tau = 0$; // simple counter for replacement

a $\tau = \left(2 \log\left(\frac{2N}{M}\right) \|a_n\|_2^2 + \|\Psi_{I_n^t} x_n\|_2^2\right)/d$; // advanced counter for adaptivity

foreach $k \in I_n^t$ **do**
$\bar{\psi}_k \leftarrow \bar{\psi}_k + \left[a_n + P(\psi_k) y_n\right] \cdot \operatorname{sign}(\langle \psi_k, y_n \rangle)$; // atom update
if $|x_n(k)|^2 \geq \tau$ **then**
$v(k) \leftarrow v(k) + 1$; // atom value update
end
end

// steps for replacement candidates
$i_n = \arg\max_\ell |\langle \gamma_\ell, a_n \rangle|$; // residual thresholding
$\bar{\gamma}_{i_n} \leftarrow \bar{\gamma}_{i_n} + a_n \cdot \operatorname{sign}(\langle \gamma_{i_n}, a_n \rangle)$; // candidate update

r $\tau_\Gamma = 2 \log(2K)/d$; // simple counter for replacement

a $\tau_\Gamma = 2 \log(\frac{2N_\Gamma}{d})/d$; // advanced counter for adaptivity

if $|\langle \gamma_{i_n}, a_n \rangle|^2 \geq \tau_\Gamma \|a_n\|_2^2$ **then**
$v_\Gamma(i_n) \leftarrow v_\Gamma(i_n) + 1$; // candidate value update
end

if $n \pmod{N_\Gamma} == 0 \wedge n < mN_\Gamma$ **then**
$\Gamma \leftarrow (\bar{\gamma}_1/\|\bar{\gamma}_1\|_2, \ldots, \bar{\gamma}_L/\|\bar{\gamma}_L\|_2)$; // candidate normalisation
$\bar{\Gamma} = 0$; // cand. iteration restart

a $v_\Gamma = 0$; // skip for replacement

end

a // steps for estimating sparsity level, skip for replacement

a $\theta = \left(2 \log(4K) \|a_n\|_2^2 + \|\Psi_{I_n^t} x_n\|_2^2\right)/d$;

a $\bar{S} \leftarrow \bar{S} + \sharp\{k : |x_n(k)|^2 \geq \theta\}$; // correct atoms

a $\bar{S} \leftarrow \bar{S} + \sharp\{k : |\langle \psi_k, a_n \rangle|^2 \geq \theta\}$; // missed atoms

end

$\Psi \leftarrow \left(\bar{\psi}_1/\|\bar{\psi}_1\|_2, \ldots, \bar{\psi}_K/\|\bar{\psi}_K\|_2\right)$; // atom normalisation

a $\bar{S} \leftarrow \lfloor \bar{S}/N \rceil$; // average sparsity level, skip for replacement

a/r **Output:** $\Psi, v, \Gamma, v_\Gamma, \bar{S}$; // estimated sparsity only for adaptivity

Algorithm C.2: Replacing coherent atoms (**d**elete, **m**erge, **a**dd)

Input: $\Psi, v, \Gamma, v_\Gamma, \mu_{\max}$; `// dict., score, cand., cand.score, threshold`

Reorder Γ, v_Γ s.t. $v_\Gamma(1) \geq v_\Gamma(2) \geq \ldots \geq v_\Gamma(L)$;
Find: $(k, k') = \arg\max_{i<j} |\langle \psi_i, \psi_j \rangle|$; `// most coherent atom pair`

while $|\langle \psi_k, \psi_{k'} \rangle| > \mu_{\max} \wedge \Gamma \neq []$ **do**

$h = \text{sign}(\langle \psi_k, \psi_{k'} \rangle)$;

d $\psi_{\bar{k}} = \lfloor \frac{v(k')}{v(k')+v(k)} \rceil \psi_{k'} + h \lfloor \frac{v(k)}{v(k')+v(k)} \rceil \psi_k$; `// delete`

m $\psi_{\bar{k}} = v(k')\psi_{k'} + hv(k)\psi_k$; `// or merge`

a $\psi_{\bar{k}} = \psi_{k'} + h\psi_{k'}$; `// or add`

$\psi_{\bar{k}} \leftarrow \psi_{\bar{k}} / \|\psi_{\bar{k}}\|_2$;
Find $\Lambda = \{\ell : \mu_\ell := \max_{i \neq k,k'} |\langle \gamma_\ell, \psi_i \rangle| > |\langle \psi_k, \psi_{k'} \rangle|\}$;
$\Gamma_\Lambda \leftarrow [], v_\Gamma(\Lambda) \leftarrow []$; `// discard coherent candidates`

if $\Gamma \neq []$ **then**
$\psi_k \leftarrow \psi_{\bar{k}}$; `// replace with merged atom`
$v(k) \leftarrow v(k) + v(k')$; `// update score of merged atom`
$\psi_{k'} \leftarrow \gamma_1$; `// replace with most useful candidate`
if $\mu_1 < \mu_{max}$ **then**
$v(k') = v_\Gamma(1)$; `// update with candidate score`
else
$v(k') = 0$; `// preferably replaced again`
end
$\gamma_1 \leftarrow [], v_\Gamma(1) \leftarrow []$; `// discard used candidate`
end
Find: $(k, k') = \arg\max_{i<j} |\langle \psi_i, \psi_j \rangle|$; `// update most coherent atom pair`

end

Output: $\Psi, v, \Gamma, v_\Gamma$

Algorithm C.3: Pruning coherent atoms (**m**erge)

Input: $\Psi, V = (v_1, \ldots, v_K), \mu_{\max}$; `// dictionary, last` m `scores, threshold`

$\Delta = \emptyset$; `// initialise set of atoms to delete`
$H = \Psi^\star \Psi - I_K$; `// hollow Gram matrix in absolute`
Find: $(k, k') = \arg\max_{i<j} |H(i,j)|$; `// most coherent atom pair`

while $|H(k, k')| > \mu_{\max}$ **do**

m $\psi_k \leftarrow v_{k'}(1)\psi_{k'} + \text{sign}(\langle \psi_k, \psi_{k'} \rangle) v_k(1)\psi_k$; `// merge according to most recent score`
$\psi_k \leftarrow \psi_k / \|\psi_k\|_2$;
$v_k(1) \leftarrow v_{k'}(1) + v_k(1)$; `// update most recent score`
$\Delta \leftarrow \Delta \cup \{k'\}$;
$H(k, \cdot) \leftarrow 0, H(k', \cdot) \leftarrow 0, H(\cdot, k) \leftarrow 0, H(\cdot, k') \leftarrow 0$; `// update hollow Gram matrix`
Find: $(k, k') = \arg\max_{j<i} |H(i,j)|$; `// update most coherent atom pair`

end
$\Psi_\Delta \leftarrow [], V_\Delta \leftarrow []$; `// delete atoms`

Output: Ψ, V

Algorithm C.4: Pruning unused atoms

Input: $\Psi, V = (v_1, \ldots, v_K), M, \delta$; // dictionary, last m scores, threshold, // maximally pruned atoms

foreach k **do**
  $\hat{v}(k) = \max_i v_k(i)$; // maximum of last m scores
end
$\Delta = \{k : \hat{v}(k) < M\}$; // atoms with max.scores below threshold
if $|\Delta| > \delta$ **then**
  Sort: $\hat{v}(i_1) \leq \hat{v}(i_2) \leq \cdots \leq \hat{v}(i_K)$; // sort max.score
  $\Delta = \{i_1 \ldots i_\delta\})$; // δ atoms with smallest max.scores
end
$\Psi_\Delta \leftarrow [], V_\Delta \leftarrow []$; // delete atoms

Output: Ψ, V

Algorithm C.5: Adding atoms

Input: $\Psi, V, \Gamma, v_\Gamma, \mu_{\max}, M$

Sort: $v_\Gamma(i_1) \geq v_\Gamma(i_2) \geq \ldots \geq v_\Gamma(i_L)$; // sort according to score
$\bar{L} = |\{\ell : v_\Gamma(\ell) \geq d\}|$; // score above d
for $\ell = 1 \ldots \bar{L}$ **do**
  if $\max_k |\langle \psi_k, \gamma_{i_\ell} \rangle| \leq \mu_{\max}$ **then**
    $\Psi \leftarrow (\Psi, \gamma_{i_\ell})$; // in order of score add if incoherent
    $V \leftarrow (V, M \cdot \mathbf{1})$; // set last m scores of added atoms to M
  end
end

Output: Ψ, V

References

[1] A. Agarwal, A. Anandkumar, P. Jain, P. Netrapalli, and R. Tandon. Learning sparsely used overcomplete dictionaries via alternating minimization. In *COLT 2014 (arXiv:1310.7991)*, 2014.

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

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

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

[8] 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.

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

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

[11] 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.

[12] 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.

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

[14] D.J. Field and B.A. Olshausen. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

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

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

[17] R. Gribonval and K. Schnass. Dictionary identifiability - sparse matrix-factorisation via l_1-minimisation. *IEEE Transactions on Information Theory*, 56(7):3523–3539, July 2010.

[18] R. Gribonval, R. Jenatton, and F. Bach. Sparse and spurious: dictionary learning with noise and outliers. *IEEE Transactions on Information Theory*, 61(11):6298–6319, 2015.

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

[20] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830. URL `https://www.tandfonline.com/doi/abs/10.1080/01621459.1963.10500830`.

[21] P. Irofti. The effect of atom replacement strategies on dictionary learning. In *iTWIST*, 2016.

[22] K. Kreutz-Delgado and B.D. Rao. FOCUSS-based dictionary learning algorithms. In *SPIE 4119*, 2000.

[23] 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.

[24] R. Kueng and D. Gross. RIPless compressed sensing from anisotropic measurements. *Linear Algebra and its Applications*, 441:110–123, 2014.

[25] M. Ledoux and M. Talagrand. *Probability in Banach spaces. Isoperimetry and processes.* Springer-Verlag, Berlin, Heidelberg, NewYork, 1991. ISBN 3-540-52013-9.

[26] M. S. Lewicki and T. J. Sejnowski. Learning overcomplete representations. *Neural Computations*, 12(2):337–365, 2000.

[27] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Discriminative learned dictionaries for local image analysis. IMA Preprint Series 2212, University of Minnesota, 2008.

[28] 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.

[29] J. Mairal, F. Bach, and J. Ponce. Task-driven dictionary learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(4):791–804, 2012.

[30] V. Naumova and K. Schnass. Fast dictionary learning from incomplete data. *EURASIP Journal on Advances in Signal Processing*, 2018(12), 2018.

[31] M. Nazzal, F. Yeganli, and H. Ozkaramanli. Improved dictionary learning by constrained re-training over residual components. In *24th Signal Processing and Communication Application Conference (SIU)*, 2016.

[32] M.C. Pali. *Dictionary Learning & Sparse Modelling.* PhD thesis, University of Innsbruck, 2021 (submitted).

[33] 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.

[34] 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.

[35] Q. Qu, Y. Zhai, X. Li, Y. Zhang, and Z. Zhu. Analysis of the optimization landscapes for overcomplete representation learning analysis of the optimization landscapes for overcomplete representation learning. *arXiv:1912.02427*, 2019.

[36] R. Rubinstein, A. Bruckstein, and M. Elad. Dictionaries for sparse representation modeling. *Proceedings of the IEEE*, 98(6):1045–1057, 2010.

[37] S. Ruetz and K. Schnass. Submatrices with non-uniformly selected random supports and insights into sparse approximation. *arXiv:2012.02082*, 2020.

[38] C. Rusu and B. Dumitrescu. Stagewise K-SVD to design efficient dictionaries for sparse representations. *IEEE Signal Processing Letters*, 19(10):631–634, 2012.

[39] C. Rusu and K. Schnass. The adaptive dictionary learning toolbox (extended abstract). In *SPARS19*, 2019.

[40] 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.

[41] K. Schnass. Local identification of overcomplete dictionaries. *Journal of Machine Learning Research (arXiv:1401.6354)*, 16(Jun):1211–1242, 2015.

[42] K. Schnass. A personal introduction to theoretical dictionary learning. *Internationale Mathematische Nachrichten*, 228:5–15, 2015.

[43] K. Schnass. Convergence radius and sample complexity of ITKM algorithms for dictionary learning. *Applied and Computational Harmonic Analysis*, 45(1):22–58, 2018.

[44] K. Schnass. Dictionary learning - from local towards global and adaptive. *arXiv:1804.07101v1*, 2018.

[45] K. Skretting and K. Engan. Recursive least squares dictionary learning algorithm. *IEEE Transactions on Signal Processing*, 58(4):2121–2130, 2010.

[46] D. Spielman, H. Wang, and J. Wright. Exact recovery of sparsely-used dictionaries. In *COLT 2012 (arXiv:1206.5882)*, 2012.

[47] J. Sun, Q. Qu, and J. Wright. Complete dictionary recovery over the sphere I: Overview and geometric picture. *IEEE Transactions on Information Theory*, 63(2):853–884, 2017.

[48] J. Sun, Q. Qu, and J. Wright. Complete dictionary recovery over the sphere II: Recovery by Riemannian trust-region method. *IEEE Transactions on Information Theory*, 63 (2):885–915, 2017.

[49] J. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

[50] J.A. Tropp. On the conditioning of random subdictionaries. *Applied and Computational Harmonic Analysis*, 25(1):1–24, 2008.