Convergence radius and sample complexity of ITKM algorithms for dictionary learning

Karin Schnass

Abstract

In this work we show that iterative thresholding and K means (ITKM) algorithms can recover a generating dictionary with K atoms from noisy S sparse signals up to an error $\tilde{\varepsilon}$ as long as the initialisation is within a convergence radius, that is up to a $\log K$ factor inversely proportional to the dynamic range of the signals, and the sample size is proportional to $K \log K \tilde{\varepsilon}^{-2}$. The results are valid for arbitrary target errors if the sparsity level is of the order of the square root of the signal dimension d and for target errors down to $K^{-\ell}$ if S scales as $S \leq d/(\ell \log K)$.

Index Terms

dictionary learning, sparse coding, sparse component analysis, sample complexity, convergence radius, alternating optimisation, thresholding, K-means

1 Introduction

The goal of dictionary learning is to find a dictionary that will sparsely represent a class of signals. That is given a set of N training signals $y_n \in \mathbb{R}^d$, which are stored as columns in a matrix $Y = (y_1, \ldots, y_N)$, one wants to find a collection of K normalised vectors $\phi_k \in \mathbb{R}^d$, called atoms, which are stored as columns in the dictionary matrix $\Phi = (\phi_1, \ldots, \phi_K) \in \mathbb{R}^{d \times K}$, and coefficients x_n, which are stored as columns in the coefficient matrix $X = (x_1, \ldots, x_N)$ such that

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

Research into dictionary learning comes in two flavours corresponding to the two origins of the problem, the slightly older one in the independent component analysis (ICA) and blind source separation (BSS) community, where dictionary learning is also known as sparse component analysis, and the slightly younger one in the signal processing community, where it is also known as sparse coding. The main motivation for dictionary learning in the ICA/BSS community comes from the assumption that the signals of interest are generated as sparse mixtures - random sparse mixing coefficients X_0 - of several sources or independent components - the dictionary Φ_0 - which can be used to describe or explain a (natural) phenomenon, [15], [30], [27], [26]. For instance in the 1996 paper by Olshausen and Field, [15], which is widely regarded as the mother contribution to dictionary learning, the dictionary is learned on patches of natural images, and the resulting atoms bear a striking similarity to simple cell receptive fields in the visual cortex. A natural question in this context is, when the generating dictionary Φ_0 can be identified from Y, that is, the sources from the mixtures. Therefore the first theoretical insights into dictionary learning came from this community, [18]. Also the first dictionary recovery algorithms with global success guarantees, which are based on finding overlapping clusters in a graph derived from the signal correlation matrix $Y^\star Y$, take the ICA/BSS point of view, [6], [2].

The main motivation for dictionary learning in the signal processing community is that sparse signals are immensely practical, as they can be easily stored, denoised, or reconstructed from incomplete information, [13], [33], [31]. Thus the interest is less in the dictionary itself but in the fact that it will provide sparse

Karin Schnass is with the Department of Mathematics, University of Innsbruck, Technikerstraße 13, 6020 Innsbruck, Austria, karin.schnass@uibk.ac.at.

representations X. Following the rule 'the sparser - the better' the obvious next step is to look for the dictionary that provides the sparsest representations. So given a budget of K atoms and S non-zero coefficients per signal, one way to concretise the abstract formulation of the dictionary learning problem in (1) is to formulate it as optimisation problem, such as

$$(P_{2,S}) \qquad \min \|Y - \Phi X\|_F \quad \text{s.t.} \quad \|x_n\|_0 \leq S \quad \text{and} \quad \Phi \in \mathcal{D}, \tag{2}$$

where $\|\cdot\|_0$ counts the nonzero elements of a vector or matrix and $\mathcal{D}$ is defined as $\mathcal{D} = \{\Phi = (\phi_1, \ldots, \phi_K) : \|\phi_k\|_2 = 1\}$. While $(P_{2,S})$ is for instance the starting point for the MOD or K-SVD algorithms, [14], [3], other definitions of *optimally* sparse lead to other optimisation problems and algorithms, [49], [37], [48], [32], [43], [38]. The main challenge of optimisation programmes for dictionary learning is finding the global optimum, which is hard because the constraint manifold $\mathcal{D}$ is not convex and the objective function is invariant under sign changes and permutations of the dictionary atoms with corresponding sign changes and permutations of the coefficient rows. In other words for every local optimum there are $2^K K! - 1$ equivalent local optima.

So while in the signal processing setting there is a priori no concept of a generating dictionary, it is often used as auxiliary assumption to get theoretical insights into the optimisation problem. Indeed without the assumption that the signals are sparse in some dictionary the optimisation formulation makes little or no sense. For instance if the signals are uniformly distributed on the sphere in $\mathbb{R}^d$, in asymptotics $(P_{2,S})$ becomes a covering problem and the set of optima is invariant under orthonormal transforms.

Based on a generating model on the other hand it is possible to gain several theoretical insights. For instance, how many training signals are necessary such that the sparse representation properties of a dictionary on the training samples (e.g. the optimiser) will extrapolate to the whole class, [34], [47], [35], [20]. What are the properties of a generating dictionary and the maximum sparsity level of the coefficients and signal noise such that this dictionary is a local optimiser or near a local optimiser given enough training signals, [21], [17], [39], [40], [19].

An open problem for overcomplete dictionaries with some first results for bases, [44], [45], is whether there are any spurious optimisers which are not equivalent to the generating dictionary, or if any starting point of a descent algorithm will lead to a global optimum? A related question (in case there are spurious optima) is, if the generating dictionary is the global optimiser? If yes, it would justify using one of the graph clustering algorithms for recovering the optimum, [6], [2], [4], [7]. This is important since all dictionary learning algorithms with global success guarantees are computationally very costly, while optimisation approaches are locally very efficient and robust to noise. Knowledge of the convergence properties of a descent algorithm, such as convergence radius (basin of attraction), rate or limiting precision based on the number of training signals, therefore helps to decide when it should take over from a global algorithm for fast local refinement, [1].

In this paper we will investigate the convergence properties of two iterative thresholding and K-means algorithms. The first algorithm ITKsM, which uses signed signal means, originates from the response maximisation principle introduced in [40]. There it is shown that a generating μ-coherent dictionary constitutes a local maximum of the response principle as long as the sparsity level of the signals scales as $S = O(\mu^{-1})$. It further contains the first results showing that the maximiser remains close to the generator for sparsity levels up to $S = O(\mu^{-2}/\log K)$. For a target recovery error $\tilde{\varepsilon}$ the sample complexity N is shown to scale as $N = O(SK^3\tilde{\varepsilon}^{-2})$ and the basin of attraction is conjectured to be of size $O(1/\sqrt{S})$.

Here we will not only improve on the conjecture by showing that in its online version the algorithm has a convergence radius of size $O(1/\sqrt{\log K})$ but also show that for the algorithm rather than the principle the sample complexity reduces to $N = O(K \log K \tilde{\varepsilon}^{-2} \log(\varepsilon^{-1}))$ (omitting $\log\log$ factors). Again recovery to arbitrary precision holds for sparsity levels $S = O(\mu^{-1})$ and stable recovery up to an error $K^{-\ell}$ for sparsity levels $S = O(\mu^{-2}/(\ell \log K))$. We also show that the computational complexity assuming an initialisation within the convergence radius scales as $O(\log(\tilde{\varepsilon}^{-1})dKN)$ or omitting log factors $O(dK^2\tilde{\varepsilon}^{-2})$.

Motivated by the desire to reduce the sample complexity for the case of exactly sparse, noiseless signals, we then introduce a second iterative thresholding and K-means algorithms ITKrM, which uses residual instead of signal means. It has roughly the same properties as ITKsM apart from the convergence radius which reduces to $O(1/\sqrt{S})$ and the computational complexity, which scales as $O(dN(K + S^2))$ and thus

can go up to $O(d^2NK)$ for $S = O(d)$. However, if $S = O(\mu^{-1})$ and the signals follow an exactly sparse, noiseless model, we can show that the sample complexity reduces to $N = O(K\tilde{\varepsilon}^{-1}\log(\varepsilon^{-1}))$ (omitting $\log\log$ factors). Our results are in the same spirit as the results for the alternating minimisation algorithm in [1] but have the advantage that they are valid for more general coefficient distributions and a lower level of sparsity (S larger) resp. higher level of coherence, that the convergence radius is larger and that the algorithms exhibit a lower computational complexity. They are also close to some very recent results about several alternating minimisation algorithms, which are like the ITKMs based on thresholding, [5]. Compared to our results they are essentially the same in terms of convergence radius and sample complexity but are only valid for sparsity levels $S = O(\mu^{-1})$ and up to a limiting precision (even in the exact sparse noiseless case). More interestingly [5] contains a strategy for finding initialisations within a radius $O(1/\log K)$ to the generating dictionary, which is proven to succeed for sparsity levels $S = O(\mu^{-1})$. With slight modifications and using Tropp's results on average isometry constants, [46], this initialisation strategy could probably be proven to work also for sparsity levels up to $S = O(\mu^{-2}/(\ell\log K))$. However, its computational complexity seems to explode as S grows.

The rest of the paper is organised as follows. After summarising notation and conventions in the following section, in Section 3 we re-introduce the ITKsM algorithm, discuss our sparse signal model and analyse the convergence properties of ITKsM. Based on the shortcomings of ITKsM we motivate the ITKrM algorithm in Section 4, and again analyse its convergence properties. In Section 5 we provide numerical simulations indicating that the convergence radius of both ITKM algorithms is generically much larger and that sometimes ITKrM even converges globally from random initialisations. Finally in Section 6 we compare our results to existing work and point out future directions of research.

2 Notations and Conventions

Before we join the melee, we collect some definitions and lose a few words on notations; usually subscripted letters will denote vectors with the exception of $\varepsilon, \alpha, \omega$, where they are numbers, eg. $x_n \in \mathbb{R}^K$ 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 $\|M\|_F = \mathrm{tr}(M^\star M)^{1/2}$, 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, i.e. $\Phi = (\phi_1, \dots \phi_K)$. The maximal absolute inner product between two different atoms is called the **coherence** μ of a dictionary, $\mu = \max_{k\neq j}|\langle\phi_k, \phi_j\rangle|$.

By Φ_I we denote the restriction of the dictionary to the atoms indexed by I, i.e. $\Phi_I = (\phi_{i_1}, \dots, \phi_{i_S})$, $i_j \in I$, and by $P(\Phi_I)$ the orthogonal projection onto the span of the atoms indexed by I, i.e. $P(\Phi_I) = \Phi_I\Phi_I^\dagger$. 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 $Q(\Phi_I)$ the orthogonal projection onto the orthogonal complement of the span on Φ_I, that is $Q(\Phi_I) = \mathbb{I}_d - P(\Phi_I)$, where $\mathbb{I}_d$ is the identity operator (matrix) in $\mathbb{R}^d$.

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

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 $d(\Psi, \Phi) = \sqrt{2 - 2/\sqrt{d}}$. A **symmetric distance** between two dictionaries Φ, Ψ could be defined as

the maximal distance between two corresponding atoms, i.e.

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

where $\mathcal{P}$ is the set of permutations of $\{1, \dots, S\}$. Since locally the distances are equivalent we will state our results 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 will make heavy use of 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

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

The dictionary Z collects the perturbation vectors on its columns, that is $Z = (z_1, \dots z_K)$ and we define the diagonal matrices A_I, W_I implicitly via

$$\Psi_I = \Phi_I A_I + Z_I W_I, \tag{6}$$

or in MATLAB notation $A_I = \mathrm{diag}(\alpha_I)$ with $\alpha_I = (\alpha_k)_{k \in I}$ and analogue for W_I. Based on this decomposition we further introduce the short hand $b_k = \frac{\omega_k}{\alpha_k} z_k$ and $B_I = Z_I W_I A_I^{-1}$.
We consider a **frame** F a collection of $K \geq d$ vectors $f_k \in \mathbb{R}^d$ for which there exist two positive constants A, B such that for all $v \in \mathbb{R}^d$ we have

$$A\|v\|_2^2 \leq \sum_{k=1}^{K} |\langle f_k, v \rangle|^2 \leq B\|v\|_2^2. \tag{7}$$

If B can be chosen equal to A, i.e. $B = A$, the frame is called tight and if all elements of a tight frame have unit norm we have $B = A = K/d$. The operator $FF^\star$ is called frame operator and by (7) its spectrum is bounded by A, B. For more details on frames, see e.g. [12].
Finally we introduce the Landau symbols O, o to characterise the growth of a function. We write

$$f(t) = O(g(t)) \quad \text{if} \quad \lim_{t \to 0/\infty} f(t)/g(t) = C < \infty$$
$$\text{and} \quad f(t) = o(g(t)) \quad \text{if} \quad \lim_{t \to 0/\infty} f(\varepsilon)/g(\varepsilon) = 0.$$

3 Dictionary Learning via ITKsM

Iterative thresholding and K signal means (ITKsM) for dictionary learning was introduced as algorithm to maximise the S-response criterion

$$(P_{R1}) \qquad \max_{\Psi \in \mathcal{D}} \sum_n \max_{|I|=S} \|\Psi_I^\star y_n\|_1, \tag{8}$$

which for $S = 1$ reduces to the K-means criterion, [40]. It belongs to the class of alternating optimisation algorithms for dictionary learning, which alternate between updating the sparse coefficients based on the current version of the dictionary and updating the dictionary based on the current version of the coefficients, [14], [3], [1]. As its name suggests, the update of the sparse coefficients is based on thresholding while the update of the dictionary is based on K signal means.

Algorithm 3.1 (ITKsM one iteration)**.** *Given an input dictionary Ψ and N training signals y_n do:*

- *For all n find $I^t_{\Psi,n} = \arg\max_{I:|I|=S} \|\Psi_I^\star y_n\|_1$.*
- *For all k calculate*

$$\bar{\psi}_k = \frac{1}{N} \sum_n y_n \cdot \mathrm{sign}(\langle \psi_k, y_n \rangle) \cdot \chi(I^t_{\Psi,n}, k). \tag{9}$$

- *Output* $\bar{\Psi} = (\bar{\psi}_1/\|\bar{\psi}_1\|_2, \dots, \bar{\psi}_K/\|\bar{\psi}_K\|_2)$.

The algorithm can be stopped after a fixed number of iterations or once a stopping criterion, such as improvement $d(\bar{\Psi}, \Psi) \leq \theta$ for some threshold θ, is reached. Its advantages over most other dictionary learning algorithms are threefold. First it has very low computational complexity. In each step the most costly operation is the calculation of the N matrix vector products $\Psi^\star y_n$, that is the matrix product $\Psi^\star Y$, of order $O(dKN)$. In comparison the globally successful graph clustering algorithms need to calculate the signal correlation matrix $Y^\star Y$, cost $O(dN^2)$.
Second due to its structure only one signal has to be processed at a time. Instead of calculating I_n^t for all n and calculating the sum, one simply calculates $I_{\Psi,n}^t$ for the signal at hand, updates all atoms $\bar{\psi}_k$ for which $k \in I_{\Psi,n}^t$ as $\bar{\psi}_k \to \bar{\psi}_k + y_n \cdot \mathrm{sign}(\langle \psi_k, y_n \rangle)$ and turns to the next signal. Once N signals have been processed one does the normalisation step and outputs $\bar{\Psi}$. Further in this online version only $(2K+1)d$ values corresponding to the input dictionary, the current version of the updated dictionary and the signal at hand, need to be stored rather than the $N \times d$ signal matrix. Parallelisation can be achieved in a similar way. Again for comparison, the graph clustering algorithms, K-SVD, [3], and the alternating minimisation algorithm in [1] need to store the whole signal resp. residual matrix as well as the dictionary.
The third advantage is that with high probability the algorithm converges locally to a generating dictionary Φ assuming that we have enough training signals and that these follow a sparse random model in Φ. In order to prove the corresponding result we next introduce our sparse signal model.

3.1 Signal Model

We employ the same signal model, which has already been used for the analyses of the S-response and K-SVD principles, [39], [40]. Given a $d \times K$ dictionary Φ, we assume that the signals are generated as,

$$y = \frac{\Phi x + r}{\sqrt{1 + \|r\|_2^2}}, \tag{10}$$

where x is drawn from a sign and permutation invariant probability distribution ν on the unit sphere $S^{K-1} \subset \mathbb{R}^K$ and $r = (r(1) \dots r(d))$ is a centred random 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 $t > 0$. We recall that a probability measure ν on the unit sphere is sign and permutation invariant, if for all measurable sets $\mathcal{X} \subseteq S^{K-1}$, for all sign sequences $\sigma \in \{-1, 1\}^d$ and all permutations p we have

$$\nu(\sigma\mathcal{X}) = \nu(\mathcal{X}), \quad \text{where} \quad \sigma\mathcal{X} := \{(\sigma(1)x(1), \dots, \sigma(K)x(d)) : x \in \mathcal{X}\} \tag{11}$$

$$\nu(p(\mathcal{X})) = \nu(\mathcal{X}), \quad \text{where} \quad p(\mathcal{X}) := \{(x(p(1)), \dots, x(p(K))) : x \in \mathcal{X}\}. \tag{12}$$

We can get a simple example of such a measure by taking a positive, non increasing sequence c, that is $c(1) \geq c(2) \geq \dots \geq c(K) \geq 0$, choosing a sign sequence σ and a permutation p uniformly at random and setting $x = x_{p,\sigma}$ with $x_{p,\sigma}(k) = \sigma(k)c(p(k))$. Conversely we can factorise any sign and permutation invariant measure into a random draw of signs and permutations and a measure on the space of non-increasing sequences.
By abuse of notation let c now denote the mapping that assigns to each $x \in S^{K-1}$ the non increasing rearrangement of the absolute values of its components, i.e. $c : x \to c_x$ with $c_x(k) := |x(p(k))|$ for a permutation p such that $|x(p(1))| \geq |x(p(2))| \geq \dots \geq |x(p(K))| \geq 0$. Then the mapping c together with the probability measure ν on $x \in S^{K-1}$ induces a probability measure ν_c on $c(S^{K-1}) = S^{K-1} \cap [0,1]^K$ via the preimage c^{-1}, that is $\nu_c(\Omega) := \nu(c^{-1}(\Omega))$ for any measurable set $\Omega \subseteq c(S^{K-1})$.
Using this new measure we can rewrite our signal model as

$$y = \frac{\Phi x_{c,p,\sigma} + r}{\sqrt{1 + \|r\|_2^2}}, \tag{13}$$

where we define $x_{c,p,\sigma}(k) = \sigma(k)c(p(k))$ for a positive, non-increasing sequence c distributed according to ν_c, a sign sequence σ and a permutation p distributed uniformly at random and r again a centred random subgaussian vector with parameter ρ. Note that we have $\mathbb{E}(\|r\|_2^2) \leq d\rho^2$, with equality for instance in the case of Gaussian noise. To incorporate sparsity into our signal model we make the following definitions.

Definition 3.1. *A sign and permutation invariant coefficient distribution ν on the unit sphere $S^{K-1} \subset \mathbb{R}^K$ is called S-sparse with absolute gap $\beta_S > 0$ and relative gap $\Delta_S > \beta_S$, if*

$$\nu\left(c_x(S) - c_x(S+1) < \beta_S\right) = 0 \qquad \textit{and} \qquad \nu\left(\frac{c_x(S) - c_x(S+1)}{c_x(1)} < \Delta_S\right) = 0, \tag{14}$$

or equivalently

$$\nu_c\left(c(S) - c(S+1) < \beta_S\right) = 0 \qquad \textit{and} \qquad \nu_c\left(\frac{c(S) - c(S+1)}{c(1)} < \Delta_S\right) = 0. \tag{15}$$

The coefficient distribution is called strongly S-sparse if $\Delta_S \geq 2\mu S$.

For exactly sparse signals β_S is simply the smallest non-zero coefficient and Δ_S is the inverse dynamic range of the non-zero coefficients. We have the bounds $\beta_S \leq \frac{1}{\sqrt{S}}$ and $\Delta_S \leq 1$. Since equality holds for the 'flat' distribution generated from $c(k) = \frac{1}{\sqrt{S}}$ for $k \leq S$ and zero else, we will usually think of β_S being of the order $O(\frac{1}{\sqrt{S}})$ and Δ_S being of the order $O(1)$. We can also see that coefficient distributions can only be strongly S-sparse as long as S is smaller than $\frac{\Delta_S}{2\mu}$, that is $S = O(\mu^{-1}) = O(\sqrt{d})$.
For the statement of our results we will use three other signal statistics,

$$\gamma_{1,S} := \mathbb{E}_c\left(c(1) + \ldots + c(S)\right) \qquad \gamma_{2,S} := \mathbb{E}_c\left(c^2(1) + \ldots + c^2(S)\right) \qquad C_r := \mathbb{E}_r\left(\frac{1}{\sqrt{1 + \|r\|_2^2}}\right). \tag{16}$$

The constants $\gamma_{1,S}$ and C_r^2 will help characterise the expected size of $\bar{\psi}_k$. We have $S\beta_S \leq \gamma_{1,S} \leq \sqrt{S}$ and

$$C_r \geq \frac{1 - e^{-d}}{\sqrt{1 + 5d\rho^2}}, \tag{17}$$

compare [40]. From the above inequality we can see that C_r captures the expected signal to noise ratio, that is for large ρ we have

$$C_r^2 \approx \frac{1}{d\rho^2} \approx \frac{\mathbb{E}(\|\Phi x\|_2^2)}{\mathbb{E}(\|r\|_2^2)}. \tag{18}$$

Similarly the constant $\gamma_{2,S}$ can be interpreted as the expected energy of the signal approximation using the largest S generating coefficients and the generating dictionary, or in other words $1 - \gamma_{2,S}$ is a bound for the expected energy of the approximation error.
For noiseless signals generated from the flat distribution described above we have $\gamma_{1,S} = \sqrt{S}$, $C_r = 1$ and $\gamma_{2,S} = 1$, so we will usually think of these constants having the orders $\gamma_{1,S} = O(\sqrt{S})$, $C_r = O(1)$ and $\gamma_{2,S} = O(1)$.
From the discussion we see that, while being relatively simple, our signal model allows us to capture both approximation error and noise. Our results have quite straightforward extensions to more complicated (realistic) signal models, which for instance include outliers (normalised but not sign or permutation invariant coefficients) or a small portion of coefficients without gap. With somewhat more effort it is also possible to relax the assumption of sign and permutation invariance in our coefficient model, potentially at the cost of decreasing the admissible sparsity level, the convergence radius and the recovery accuracy and increasing the sample complexity. Indeed we will see that the main reason for assuming sign invariance is to ensure that when thresholding with the generating dictionary always succeeds in recovering the generating support with a large margin and therefore also succeeds with a perturbed dictionary. To a lesser degree, especially in the case of ITKrM, the sign invariance also supports the permutation invariance in ensuring a richness of signals such that the averaging procedures contract towards the generating atoms. In particular the permutation invariance prevents the situation that two atoms are always used together and could therefore be replaced by two of their linear combinations.
However, we will sacrifice generality for comprehensibility and therefore just give pointers in the respective proofs.

3.2 Convergence analysis of ITKsM

We first look at the more general case of noisy, non exactly S-sparse signals and specialise to noiseless, strongly S-sparse signals later.

Theorem 3.2. *Let Φ be a unit norm frame with frame constants $A \leq B$ and coherence μ and assume that the training signals y_n are generated according to the signal model in* (13) *with coefficients that are S-sparse with absolute gap β_S and relative gap Δ_S.*
Fix a target error $\tilde{\varepsilon} \geq 4\varepsilon_{\mu,\rho}$, where

$$\varepsilon_{\mu,\rho} := \frac{8K^2\sqrt{B+1}}{C_r\gamma_{1,S}} \exp\left(\frac{-\beta_S^2}{98\max\{\mu^2,\rho^2\}}\right). \tag{19}$$

Given an input dictionary Ψ such that

$$d(\Psi,\Phi) \leq \frac{\Delta_S}{\sqrt{98B}\left(\frac{1}{4}+\sqrt{\log\left(\frac{1060K^2(B+1)}{\Delta_S C_r\gamma_{1,S}}\right)}\right)}, \tag{20}$$

then after $6\lceil\log(\tilde{\varepsilon}^{-1})\rceil$ iterations of ITKsM each on a fresh batch of N training signals the output dictionary $\tilde{\Psi}$ satisfies

$$d(\tilde{\Psi},\Phi) \leq \tilde{\varepsilon} \tag{21}$$

except with probability

$$18\lceil\log(\tilde{\varepsilon}^{-1})\rceil K \exp\left(\frac{-C_r^2\gamma_{1,S}^2 N\tilde{\varepsilon}^2}{200SK}\right).$$

Before providing the proof let us discuss the result above. We first see that ITKsM will succeed if the input dictionary is within a radius $O(\Delta_S/\sqrt{\log K})$ to the generating dictionary Φ. In case of exactly sparse signals this means that the convergence radius is up to a log factor inversely proportional to the dynamic range of the coefficients. This should not be come as a big surprise, considering that the average success of thresholding for sparse recovery with a ground truth dictionary depends on the dynamic range, [42]. It also means that in the best case the convergence radius is actually of size $O(1/\sqrt{\log K})$, since for the flat distribution $\Delta_S = 1$.
Next note that in the theorem above we have restricted the target error to be larger than $4\varepsilon_{\mu,\rho}$. However at the cost of unattractively large constants in the probability bound, we can actually reach any target error larger than $\varepsilon_{\mu,\rho}$.
To highlight the relation between the sparsity level and the minimally achievable error, we specialise the result to coefficients drawn from the flat distribution, meaning $\beta_S = 1/\sqrt{S}$. We further assume white Gaussian noise with variance $\rho^2 = 1/d$, corresponding to an expected signal to noise ratio of 1, and an incoherent dictionary with $\mu \leq 1/\sqrt{d}$. If $S \leq \frac{d}{98\ell\log K}$ for some $\ell \geq 2$ then the minimally achievable error $\varepsilon_{\mu,\rho}$ can be as small as $O(K^{2-\ell})$.
Last we want to get a feeling for the total number of training signals we need to have a good success probability. For exactly S-sparse signals with dynamic coefficient range 1 we have $\gamma_{1,S} = \sqrt{S}$. Omitting loglog factors each iteration is therefore likely to be successful when using a batch of $N = O(K\log K\tilde{\varepsilon}^{-2})$ training signals, meaning that ITKsM is successful with high probability as soon as the total number of training signals used in the algorithms scales as $O(K\log K\tilde{\varepsilon}^{-2}\log(\tilde{\varepsilon}^{-1}))$. Note that in case of noise due to information theoretic arguments the factor $\tilde{\varepsilon}^{-2}$ seems unavoidable, [25].
To summarise the discussion we provide an O-notation version of the theorem, which is less plug and play but free of messy constants and as such better suited to convey the quality of the result. Compare also Subsection 3.1 for the O notation conventions.

Theorem - O (3.2). *Assume that in each iteration the number of training signals scales as $N = O(K\log K\tilde{\varepsilon}^{-2})$. If $S \leq O(\frac{1}{\ell\mu^2\log K})$ then with high probability for any starting dictionary Ψ within distance $\varepsilon \leq O(1/\sqrt{\log K})$*

to the generating dictionary after $O(\log(\tilde{\varepsilon}^{-1}))$ iterations of ITKsM, each on a fresh batch of training signals, the distance of the output dictionary $\tilde{\Psi}$ to the generating dictionary will be smaller than

$$\max\left\{\tilde{\varepsilon}, O\left(K^{2-\ell}\right)\right\}. \tag{22}$$

Proof: The proof consists of two steps. First we show that with high probability one iteration of ITKsM reduces the error by at least a factor $\kappa < 1$. Then we iteratively apply the results for one iteration. *Step 1:* For the first step we use the following ideas, compare also [40]: For most sign sequences σ_n and therefore most signals

$$y_n = \frac{\Phi x_{c_n,p_n,\sigma_n} + r_n}{\sqrt{1+\|r_n\|_2^2}}$$

thresholding with a perturbation of the original dictionary will still recover the generating support $I_n := p_n^{-1}(\mathbb{S})$, that is $I_{\Psi,n}^t = I_n$. Assuming that the generating support is recovered, for each k the expected difference of the sum in (9) between using the original Φ and the perturbation Ψ is small, that is smaller than $d(\Phi,\Psi) = \varepsilon$, and due to concentration of measure also the difference on a finite number of samples will be small. Finally for each k the sum in (9) will again concentrate around its expectation, a scaled version of the atom ϕ_k.
Formally we write,

$$\bar{\psi}_k = \frac{1}{N}\sum_n y_n \operatorname{sign}(\langle\psi_k, y_n\rangle)\,\chi(I_{\Psi,n}^t, k) - \frac{1}{N}\sum_n y_n\,\sigma_n(k)\,\chi(I_n,k) \tag{23}$$

$$+\frac{1}{N}\sum_n y_n\,\sigma_n(k)\,\chi(I_n,k) - \mathbb{E}\left(\frac{1}{N}\sum_n y_n\,\sigma_n(k)\,\chi(I_n,k)\right) + \mathbb{E}\left(\frac{1}{N}\sum_n y_n\,\sigma_n(k)\,\chi(I_n,k)\right). \tag{24}$$

Since $\mathbb{E}\left(\frac{1}{N}\sum_n y_n\,\sigma_n(k)\,\chi(I_n,k)\right) = \frac{C_r\gamma_{1,S}}{K}\phi_k$, see the proof of Lemma B.5 in the appendix, using the triangle inequality and the bound $\|y_n\|_2 \le \sqrt{B+1}$ we get,

$$\begin{aligned}\left\|\bar{\psi}_k - \frac{C_r\gamma_{1,S}}{K}\phi_k\right\|_2 &\le \left\|\frac{1}{N}\sum_n y_n\left[\operatorname{sign}(\langle\psi_k,y_n\rangle)\,\chi(I_{\Psi,n}^t,k) - \sigma_n(k)\,\chi(I_n,k)\right]\right\|_2 \\ &\quad + \left\|\frac{1}{N}\sum_n y_n\,\sigma_n(k)\,\chi(I_n,k) - \frac{C_r\gamma_{1,S}}{K}\phi_k\right\|_2 \\ &\le \frac{2\sqrt{B+1}}{N}\,\sharp\{n : \operatorname{sign}(\langle\psi_k,y_n\rangle)\,\chi(I_{\Psi,n}^t,k) \ne \sigma_n(k)\,\chi(I_n,k)\} \\ &\quad + \left\|\frac{1}{N}\sum_n y_n\,\sigma_n(k)\,\chi(I_n,k) - \frac{C_r\gamma_{1,S}}{K}\phi_k\right\|_2. \end{aligned} \tag{25}$$

Next note that for the draw of y_n the event that for a given index k the signal coefficient using thresholding with Ψ is different from the oracle signal is contained in the event that thresholding does not recover the entire generating support $I_{\Psi,n}^t \ne I_n$ or that on the generating support the empirical sign pattern using Ψ is different from the generating pattern, $\operatorname{sign}(\langle\psi_k,y_n\rangle) \ne \sigma_n(k)$ for a $k \in I_n$,

$$\{y_n : \operatorname{sign}(\langle\psi_k,y_n\rangle)\,\chi(I_{\Psi,n}^t,k) \ne \sigma_n(k)\,\chi(I_n,k)\} \subseteq \{y_n : I_{\Psi,n}^t \ne I_n\} \cup \{y_n : \operatorname{sign}(\Psi_{I_n}^\star y_n) \ne \sigma_n(I_n)\}. \tag{26}$$

From [40], e.g. proof of Proposition 7, we know that the right hand side in (26) is in turn contained in the event $\mathcal{E}_n \cup \mathcal{F}_n$, where

$$\mathcal{E}_n := \left\{y_n : \exists k \text{ s.t. } \left|\sum_{j\ne k}\sigma_n(j)c_n\big(p_n(j)\big)\langle\phi_j,\phi_k\rangle\right| \ge u_1 \text{ or } |\langle r_n,\phi_k\rangle| \ge u_2\right\} \tag{27}$$

$$\mathcal{F}_n := \left\{y_n : \exists k \text{ s.t. } \omega_k\left|\sum_{j}\sigma_n(j)c_n\big(p_n(j)\big)\langle\phi_j,z_k\rangle\right| \ge u_3 \text{ or } \omega_k|\langle r_n,z_k\rangle| \ge u_4\right\} \tag{28}$$

$$\text{for}\quad 2(u_1+u_2+u_3+u_4) \le c_n(S)\left(1-\frac{\varepsilon^2}{2}\right) - c_n(S+1). \tag{29}$$

In particular if we choose $u_1 = u_2 = (c_n(S) - c_n(S+1))/7$, $u_3 = u_1 - \frac{\varepsilon^2 c_n(S)}{6}$ and $u_4 = u_3/2$ we get that $\mathcal{E}_n$, which contains the event that thresholding using the generating dictionary Φ fails, is independent of Ψ. To estimate the number of signals for which the thresholding summand is different from the oracle summand, it suffices to count how often $y_n \in \mathcal{E}_n$ or $y_n \in \mathcal{F}_n$,

$$\sharp\{n : \operatorname{sign}(\langle \psi_k, y_n \rangle)\, \chi(I^t_{\Psi,n}, k) \neq \sigma_n(k)\, \chi(I_n, k)\} \leq \sharp\{n : y_n \in \mathcal{E}_n\} + \sharp\{n : y_n \in \mathcal{F}_n\}. \tag{30}$$

Substituting these bounds into (25) we get,

$$\begin{aligned}\left\| \bar{\psi}_k - \frac{C_r \gamma_{1,S}}{K} \phi_k \right\|_2 \leq \frac{2\sqrt{B+1}}{N} \sharp\{n : y_n \in \mathcal{E}_n\} + \frac{2\sqrt{B+1}}{N} \sharp\{n : y_n \in \mathcal{F}_n\} \\ + \left\| \frac{1}{N} \sum_n y_n\, \sigma_n(k)\, \chi(I_n, k) - \frac{C_r \gamma_{1,S}}{K} \phi_k \right\|_2 .\end{aligned} \tag{31}$$

If we want the error between $\bar{\psi}_k / \|\bar{\psi}_k\|_2$ and ϕ_k to be of the order $\kappa\varepsilon$, we need to ensure that the right hand side of (31) is less than $\kappa\varepsilon \cdot \frac{C_r \gamma_{1,S}}{K}$.
From Lemma B.3 in the appendix we know that

$$\mathbb{P}\left(\sharp\{n : y_n \in \mathcal{E}_n\} \geq \frac{C_r \gamma_{1,S} N}{2K\sqrt{B+1}} \cdot (\varepsilon_{\mu,\rho} + t_1) \right) \leq \exp\left(\frac{-t_1^2 C_r \gamma_{1,S} N}{2K\sqrt{B+1}\,(2\varepsilon_{\mu,\rho} + t_1)} \right). \tag{32}$$

Next Lemma B.4 tells us that

$$\mathbb{P}\left(\sharp\{n : y_n \in \mathcal{F}_n\} \geq \frac{C_r \gamma_{1,S} N}{2K\sqrt{B+1}} \cdot (\tau\varepsilon + t_2) \right) \leq \exp\left(\frac{-t_2^2 C_r \gamma_{1,S} N}{2K\sqrt{B+1}\,(2\tau\varepsilon + t_2)} \right), \tag{33}$$

whenever

$$\varepsilon \leq \frac{\Delta_S}{\sqrt{98B}\left(\frac{1}{4} + \sqrt{\log\left(\frac{106K^2(B+1)}{\Delta_S C_r \gamma_{1,S} \tau} \right)} \right)}. \tag{34}$$

Finally by Lemma B.5 we have

$$\mathbb{P}\left(\left\| \frac{1}{N} \sum_n \frac{\Phi x_{c_n,p_n,\sigma_n} + r_n}{\sqrt{1 + \|r_n\|_2^2}} \cdot \sigma_n(k) \cdot \chi(I_n, k) - \frac{C_r \gamma_{1,S}}{K} \phi_k \right\|_2 \geq t_3 \frac{C_r \gamma_{1,S}}{K} \right) \leq \exp\left(\frac{-t_3^2 C_r^2 \gamma_{1,S}^2 N}{8SK} + \frac{1}{4} \right), \tag{35}$$

whenever $0 \leq t_3 \leq \frac{\sqrt{S}}{\sqrt{B}+2}$. Thus with high probability we have,

$$\left\| \bar{\psi}_k - \frac{C_r \gamma_{1,S}}{K} \phi_k \right\|_2 \leq \frac{C_r \gamma_{1,S}}{K} \left(\varepsilon_{\mu,\rho} + t_1 + \tau\varepsilon + t_2 + t_3 \right). \tag{36}$$

To be more precise if we choose a target error $\tilde{\varepsilon} \geq 4\varepsilon_{\mu,\rho}$ and set $t_1 = \tilde{\varepsilon}/10$, $t_2 = \max\{\tilde{\varepsilon}, \varepsilon\}/10$, $\tau = 1/10$ and $t_3 = \tilde{\varepsilon}/5$, then except with probability

$$\exp\left(\frac{-C_r \gamma_{1,S} N \tilde{\varepsilon}}{120K\sqrt{B+1}} \right) + \exp\left(\frac{-C_r \gamma_{1,S} N \max\{\tilde{\varepsilon}, \varepsilon\}}{60K\sqrt{B+1}} \right) + 2K \exp\left(\frac{-C_r^2 \gamma_{1,S}^2 N \tilde{\varepsilon}^2}{200SK} \right) \tag{37}$$

we have

$$\max_k \left\| \bar{\psi}_k - \frac{C_r \gamma_{1,S}}{K} \phi_k \right\|_2 \leq \frac{C_r \gamma_{1,S}}{K} \cdot \frac{3}{4} \cdot \max\{\tilde{\varepsilon}, \varepsilon\}. \tag{38}$$

By Lemma B.10 this further implies that

$$d(\bar{\Psi}, \Phi) = \max_k \left\| \frac{\bar{\psi}_k}{\|\bar{\psi}_k\|_2} - \phi_k \right\|_2 \leq 0.83 \max\{\tilde{\varepsilon}, \varepsilon\}. \tag{39}$$

Note that in case of outliers we first have to split the sum in (9) into the outliers, whose number concentrates around N the probability of being an outlier, and the inliers for which we can use the

same procedure as above, see [19] for more details. Similarly the small portion of coefficients without (sufficiently) large gap can be included in the small number of signals for which thresholding fails.
Step 2: From Step 1 we know that in each iteration the error will either be decreased by at least a factor 0.83 or if its already below $\tilde{\varepsilon}$ will stay below $\tilde{\varepsilon}$. So after L iterations each using a new batch of N signals, $d(\tilde{\Psi}, \Phi) \leq \max\{\tilde{\varepsilon}, 0.83^L d(\Psi, \Phi)\} \leq \max\{\tilde{\varepsilon}, 0.83^L\}$, except with probability

$$L\left(\exp\left(\frac{-C_r\gamma_{1,S}N\tilde{\varepsilon}}{120K\sqrt{B+1}}\right) + \exp\left(\frac{-C_r\gamma_{1,S}N\max\{\tilde{\varepsilon},\varepsilon\}}{60K\sqrt{B+1}}\right) + 2K\exp\left(\frac{-C_r^2\gamma_{1,S}^2N\tilde{\varepsilon}^2}{200SK}\right)\right) \tag{40}$$

Setting $L = 6\lceil\log(\tilde{\varepsilon}^{-1})\rceil$ and taking into account that the failure probability of each iteration is bounded by $3K\exp\left(\frac{-C_r^2\gamma_{1,S}^2N\tilde{\varepsilon}^2}{200SK}\right)$ leads to the final estimate.

□

For most desired precisions Theorem 3.2, which is valid for a quite large hyper-cube of input dictionaries and a wide range of sparsity levels, will actually be sufficient. However, for completeness we specialise the theorem above to the case of strongly S-sparse, noiseless signals and show that in this case ITKsM can achieve arbitrarily small errors, provided enough samples.

Corollary 3.3. *Let Φ be a unit norm frame with frame constants $A \leq B$ and coherence μ and assume that the training signals y_n are generated according to the signal model in* (13) *with $r = 0$ and coefficients that are strongly S-sparse with relative gap $\Delta_S > 2\mu S$. Fix a target error $\tilde{\varepsilon} \geq 0$. If for the input dictionary Ψ we have*

$$d(\Psi, \Phi) \leq \frac{\Delta_S - 2\mu S}{\sqrt{98B}\left(\frac{1}{4} + \sqrt{\log\left(\frac{1060K^2B}{(\Delta_S - 2\mu S)\gamma_{1,S}}\right)}\right)}, \tag{41}$$

then after $6\lceil\log(\tilde{\varepsilon}^{-1})\rceil$ iterations of ITKsM, each on a fresh batch of N training signals, the output dictionary $\tilde{\Psi}$ satisfies

$$d(\tilde{\Psi}, \Phi) \leq \tilde{\varepsilon} \tag{42}$$

except with probability

$$18\log(\tilde{\varepsilon}^{-1})K\exp\left(\frac{-\gamma_{1,S}^2N\tilde{\varepsilon}^2}{200SK}\right).$$

The proof is analogue to the one of Theorem 3.2 and can be found in Appendix A.1.

Let us again discuss the result. The main difference to Theorem 3.2 is that the condition $\Delta_S \geq 2\mu S$ can only hold for much lower sparsity levels, that is $S = O(\mu^{-1})$ and thus for incoherent dictionaries up to the square root of the ambient dimension $O(\sqrt{d}) \ll O(d/\log K)$. It is also no surprise that once the input dictionary is up to a log factor within this radius, ITKsM can achieve arbitrarily small errors. Indeed once $\Delta_S \geq 2\mu S$ thresholding is always guaranteed to recover the sparse support of a signal given the ground truth dictionary or a slight perturbation of it, [42].
To again turn the corollary into something less technical and more interesting we combine it with the corresponding theorem. If the coefficients are strongly S-sparse the minimally achievable error using Theorem 3.2 will be smaller than the error we need for Corollary 3.3 to take over and so we get the following O notation result.

Corollary - O (3.3). *Assume that in each iteration the number of noiseless, exactly S-sparse training signals scales as $O(K\log K\tilde{\varepsilon}^{-2})$. If $S \leq O(\mu^{-1})$ then with high probability for any starting dictionary Ψ within distance $\varepsilon \leq O(1/\sqrt{\log K})$ to the generating dictionary after $O(\log(\tilde{\varepsilon}^{-1}))$ iterations of ITKsM, each on a fresh batch of training signals, the distance of the output dictionary $\tilde{\Psi}$ to the generating dictionary will be smaller than $\tilde{\varepsilon}$.*

While a convergence radius of around $1/\sqrt{\log K}$, admissible sparsity levels up to $d/\log K$ and a dependence of the sample complexity on only $K\log K$ is very positive, the dependence of the sample complexity on the squared inverse target error $\tilde{\varepsilon}^{-2}$ for noiseless exactly S-sparse signals is somewhat

disappointing. Again note that in the case of noisy signals information theoretic arguments indicate that this factor is unavoidable, [25]. Looking at the proof of Theorem 3.2 we see that the reason for this factor is the slow concentration of the sums $\frac{1}{N}\sum_n y_n\,\sigma_n(k)\,\chi(I_n,k)$ around the atom ϕ_k. This can in turn be explained by the fact that via the summation we have to cancel out the equally sized contribution of all other atoms. Actively trying to cancel out these contributions already before the summation, that is summing residuals instead of signals, should therefore accelerate the concentration, and lead to a lower sample complexity in case of noiseless signals and better constants in case of noisy signals. We will concretise these ideas in the next section.

4 Dictionary Learning via ITKrM

There are several ways to remove the contribution of all atoms in the current support $I^t_{\Psi,n}$ except for ψ_k. The maybe most obvious way is to consider $Q(\Psi_{I^t_{\Psi,n}\setminus k})y_n = [\mathbb{I}_d - P(\Psi_{I^t_{\Psi,n}\setminus k})]y_n$. Unfortunately this residual has several disadvantages, the most severe being that it is not clear whether for the oracle supports and oracle signs the corresponding sum of residuals concentrates around a multiple of the atom ϕ_k,

$$\mathbb{E}\left(\frac{1}{N}\sum_n Q(\Phi_{I_n\setminus k})y_n\cdot\sigma_n(k)\cdot\chi(I_n,k)\right)\propto\mathbb{E}_{I:k\in I}\left(Q(\Phi_{I\setminus k})\,\phi_k\right)\overset{?}{\propto}\phi_k. \tag{43}$$

We suspect that equality can only hold for tight dictionaries and that an additional constraint such as minimal incoherence is needed. We therefore choose a perhaps less obvious but more stable residual $a_{n,k}(\Psi) = y_n - P(\Psi_{I^t_{\Psi,n}})y_n + P(\psi_k)y_n$, which captures the contribution of the current atom ϕ_k as well as the approximation error in Ψ, that is $y_n - P(\Psi_{I^t_{\Psi,n}})y_n$. Replacing the signal means in ITKsM with residual means we arrive at the new algorithm, iterative thresholding and K residual means (ITKrM).

Algorithm 4.1 (ITKrM one iteration)**.** *Given an input dictionary Ψ and N training signals y_n do:*

- *For all n find $I^t_{\Psi,n} = \arg\max_{I:|I|=S}\|\Psi_I^\star y_n\|_1$.*
- *For all k calculate*

$$\bar\psi_k = \frac{1}{N}\sum_n \left[y_n - P(\Psi_{I^t_{\Psi,n}})y_n + P(\psi_k)y_n\right]\cdot\operatorname{sign}(\langle\psi_k,y_n\rangle)\cdot\chi(I^t_{\Psi,n},k). \tag{44}$$

- *Output $\bar\Psi = (\bar\psi_1/\|\bar\psi_1\|_2,\dots,\bar\psi_K/\|\bar\psi_K\|_2)$.*

Again ITKrM inherits most computational properties of ITKsM. As such it can again be stopped after a fixed number of iterations or once a stopping criterion, such as the improvement below some threshold, is reached. Only one signal has to be processed at a time, making it suitable for an online version and parallelisation. Its computational complexity is slightly larger than for ITKsM because of the projections $P(\Psi_{I^t_{\Psi,n}})y_n$. If computed with maximal numerical stability, these have an overall cost of $O(S^2dN)$, which corresponds to the QR decompositions of $\Psi_{I^s_n}$. However, since the achievable precision in the learning is usually 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 via the eigenvalue decompositions of $\Psi^\star_{I^s_n}\Psi_{I^s_n}$, which is less stable but reduces the overall cost to $O(S^3N)$. Still for $S \geq d^{2/3}$ these computations become the determining factor; we will see that S can again be of the order $O(\mu^{-2}/\log K) \approx O(d/\log K)$. In the next subsection we will analyse which convergence properties of ITKsM translate to ITKrM.

4.1 Convergence Analysis of ITKrM

As for ITKsM we focus on the more realistic case of non exactly S-sparse and/or relatively noisy signals and specialise our results to exactly S-sparse, noiseless signals and moreover the case where $S \leq O(\mu^{-1})$ later.

Theorem 4.2. *Let Φ be a unit norm frame with frame constants $A \leq B$ and coherence μ and assume that the training signals y_n are generated according to the signal model in* (13) *with coefficients that are S-sparse with*

absolute gap β_S and relative gap Δ_S. Assume further that $S \leq \frac{K}{98B}$ and $\varepsilon_\delta := K \exp\left(-\frac{1}{4741\mu^2 S}\right) \leq \frac{1}{48(B+1)}$. Fix a target error $\tilde{\varepsilon} \geq 8\varepsilon_{\mu,\rho}$, with

$$\varepsilon_{\mu,\rho} = \frac{8K^2\sqrt{B+1}}{C_r\gamma_{1,S}} \exp\left(\frac{-\beta_S^2}{98\max\{\mu^2,\rho^2\}}\right), \tag{45}$$

compare (19)*, and assume that $\tilde{\varepsilon} \leq 1 - \gamma_{2,S} + d\rho^2$.*
If for the input dictionary Ψ we have

$$d(\Psi,\Phi) \leq \frac{\Delta_S}{\sqrt{98B}\left(\frac{1}{4} + \sqrt{\log\left(\frac{2544K^2(B+1)}{\Delta_S C_r \gamma_{1,S}}\right)}\right)} \qquad \text{and} \qquad d(\Psi,\Phi) \leq \frac{1}{32\sqrt{S}}, \tag{46}$$

then after $12\lceil\log(\tilde{\varepsilon}^{-1})\rceil$ iterations of ITKrM each on a fresh batch of N training signals the output dictionary $\tilde{\Psi}$ satisfies

$$d(\tilde{\Psi},\Phi) \leq \tilde{\varepsilon} \tag{47}$$

except with probability

$$60\lceil\log(\tilde{\varepsilon}^{-1})\rceil K \exp\left(\frac{-C_r^2\gamma_{1,S}^2 N\tilde{\varepsilon}^2}{576K\max\{S,B+1\}\left(\tilde{\varepsilon}+1-\gamma_{2,S}+d\rho^2\right)}\right). \tag{48}$$

Proof: The proof follows the same two step procedure as the proof of Theorem 3.2, where in the first step we prove that one iteration will reduce the error by a factor $\kappa < 1$ with high probability and then iterate this result. To prove the first step we again use a triangular inequality argument. So we check how often thresholding with Ψ fails. Assuming thresholding recovers the generating support we show that the difference between the oracle residual (based on the generating sign and support) using Φ and the oracle residual using Ψ concentrates around its expectation, which is small. Finally we show that the sum of residuals using Φ converges to a scaled version of ϕ_k. To keep the flow of the paper we do not give the full proof here but in Appendix A.2. □

Let us discuss the result. First we see that compared to the corresponding theorem for ITKsM we need somewhat more conditions. The first two extra conditions on the sparsity level $S \leq \frac{K}{98B}$ and $48(B+1)\varepsilon_\delta < 1$ are technicalities. For all but the most ideal cases they are already implied by having a limiting error $\varepsilon_{\mu,\rho}$ smaller than one. Since $\beta_S \leq 1/\sqrt{S}$ the first condition is implied as soon as μ^2 is larger than B/K, where at best we have $\mu^2 = \frac{B-1}{K-1}$. The second condition is a substitute for having small isometry constant of the generating dictionary $\delta_S \leq \frac{1}{4}$ and guarantees that most support sets of size S have $\delta_I(\Phi) \leq \frac{1}{4}$. It is implied by $\varepsilon_{\mu,\rho} \leq 1$ as soon as β_S is smaller than $\frac{1}{7\sqrt{S}}$ or equivalently the dynamic range of the coefficients is larger than 7.
The target error can again be chosen closer to the limiting error at the cost of horrible constants. Also note that the condition that the target error should be smaller than the expected squared approximation error and noise is again a technicality. If both noise and approximation error are so small that a larger target error makes sense we get the same result but with a smaller failure probability. To get an idea how such a result would look like we refer the reader to the corollary below, where we assume exactly sparse noiseless signals.
The only extra condition that changes the quality of the result is the second condition on the convergence radius. Assuming that $\Delta_S = O(1)$ the first bound in (46) is of the order $O(1/\sqrt{\log K})$, so as soon as $S \geq \log K$, meaning for most practically relevant cases, the second bound will be more restrictive. This decreased convergence radius of ITKrM compared to ITKsM is a little disappointing but seems unavoidable. The reason for this is that the expected difference between the oracle residuals using Ψ and Φ depends on the operator norms of the rescaled perturbation matrices $\|B_I\|_{2,2}$, compare Lemma B.8. If the perturbation dictionary is quasi constant, that is before normalisation $z_k = v - P(\phi_k)v$ for some $v \neq 0$, then $\|B_I\|_{2,2} \approx \sqrt{S}\varepsilon$ for all possible subsets I, so we need $\varepsilon \leq 1/\sqrt{S}$.
The advantage over ITKrM is that for low expected noise levels and approximation errors, $1-\gamma_{2,S}+d\rho^2 \ll 1$, we get better constants in the sample complexity. Actually from the probability bound in (48) we can

already guess that for exactly sparse, noiseless signals we can reduce the factor ε^{-2} in the exponent to ε^{-1}. Before specialising the theorem to noiseless signals we again provide a qualitative result, which combines the theorem above with the corresponding theorem for ITKsM in order to deal with the reduced convergence radius. That is we first exploit the large convergence radius of ITKsM and run ITKsM to arrive at an error $O(1/\sqrt{S})$. Then we exploit the lower sample complexity of ITKrM to arrive at the target error.

Theorem - O (4.2). *Assume that in each iteration the number of training samples N scales as $O(K \log K \tilde{\varepsilon}^{-2})$. If $S \leq \frac{1}{\mu^2 \ell \log K}$ then with high probability for any starting dictionary Ψ within distance $\varepsilon \leq O(1/\sqrt{\log K})$ to the generating dictionary after $O(\log(S))$ iterations of ITKsM and $O(\log(\tilde{\varepsilon}^{-1}))$ iterations of ITKrM the distance of the output dictionary $\tilde{\Psi}$ to the generating dictionary will be smaller than*

$$\max\left\{\tilde{\varepsilon}, O\left(K^{2-\ell}\right)\right\}. \tag{49}$$

Unfortunately the better constant in the sample complexity of ITKrM disappears in the O notation and we cannot really see the improvement over ITKsM. We therefore specialise again to noiseless, strongly S-sparse signals.

Corollary 4.3. *Let Φ be a unit norm frame with frame constants $A \leq B$ and coherence μ and assume that the training signals y_n are generated according to the signal model in* (13) *with $r = 0$ and coefficients that are exactly and strongly S-sparse with relative gap $\Delta_S > 2\mu S$. Fix a target precision $\tilde{\varepsilon} > 0$. If for the input dictionary Ψ we have $d(\Psi, \Phi) \leq \frac{1}{32\sqrt{S}}$ and*

$$d(\Psi, \Phi) \leq \frac{\Delta_S - 2\mu S}{\sqrt{12}\left(\frac{1}{4} + \sqrt{\log\left(\frac{23K^2\sqrt{B}}{(\Delta_S - 2\mu S)\gamma_{1,S}}\right)}\right)} \qquad \text{and} \qquad d(\Psi, \Phi) \leq \frac{1}{32\sqrt{S}}, \tag{50}$$

then after $9\lceil \log(\tilde{\varepsilon}^{-1})\rceil$ iterations of ITKrM, each on a fresh batch of N training signals, the output dictionary $\tilde{\Psi}$ satisfies

$$d(\tilde{\Psi}, \Phi) \leq \tilde{\varepsilon}$$

except with probability

$$27K\lceil \log(\tilde{\varepsilon}^{-1})\rceil \exp\left(\frac{-\gamma_{1,S}^2 N\tilde{\varepsilon}}{144\, K \max\{S, B\}}\right). \tag{51}$$

The proof sketch can be found in the Appendix A.3.

The above corollary clearly reveals the influence of the underlying signal model on dictionary learning results. So assuming that the signals are noiseless and exactly sparse and that S is only of the order $O(\mu^{-1}) = O(\sqrt{d})$, we get that one iteration of ITKrM will reduce the error as long as the number of samples scales as $O(K\varepsilon^{-1})$, meaning the influence of the target error is reduced by a factor ε^{-1}!
Again combining with ITKsM and assuming that the stronger restriction on the convergence radius is the second bound in (50), we get the following quantitative results.

Corollary - O (4.3). *Assume that in each iteration the number of noiseless, exactly S-sparse training signals scales as $O(K \log K \tilde{\varepsilon}^{-1})$. If $S \leq O(\mu^{-1})$ then with high probability for any starting dictionary Ψ within distance $\varepsilon \leq O(1/\sqrt{\log K})$ to the generating dictionary after $O(\log(S))$ iterations of ITKsM and $O(\log(\tilde{\varepsilon}^{-1}))$ iterations of ITKrM, each on a fresh batch of training signals, the distance of the output dictionary $\tilde{\Psi}$ to the generating dictionary will be smaller than $\tilde{\varepsilon}$.*

Before a final discussion of our results we first illustrate our theoretical findings with some numerical simulations, which give interesting insights into the average convergance radius of the algorithms and indicate that in practice ITKrM can be a very powerful low complexity alternative to K-SVD.

5 NUMERICAL SIMULATIONS

To complement our theoretical findings, we conduct two small numerical experiments both on synthetic and real data[1]. First we test the average case convergence radius and speed of the ITKsM and ITKrM algorithm, by running both algorithms on noiseless and noisy training data, using three different types of initialisations with varying distance to the generating dictionary.

We generate our training signals based on the signal model in (13). As generating dictionary Φ we choose the dictionary consisting of the Dirac basis and the first half of the elements of the discrete cosine transform basis in $\mathbb{R}^d$ with $d = 256$, meaning $K = 3 * d/2 = 384$, which has coherence $\mu = \sqrt{2/d} \approx 0.088$. Given a sparsity level S, to simulate noiseless, exactly sparse signals we choose c with $c_1 = \ldots = c_S = 1/\sqrt{S}$ and $c_k = 0$ for $k > S$, meaning dynamic range 1. To simulate noisy signals with a higher dynamic range we choose a decay parameter c_b uniformely at random in [0.9,1], and let the first S entries be a geometric sequence, that is $c_k = b_0 * c_b^k$ for $k \leq S$ and $c_k = 0$ for $k > S$, where b_0 is a scaling parameter ensuring that $\|c\|_2 = 1$. The noise r is chosen as a centered Gaussian with variance $1/d$, that is $r(k) \sim \mathcal{N}(0, 1/\sqrt{d})$, resulting in an expected signal to noise ratio of 1. The three different types of initialisations are created by first choosing vectors z_k uniformly at random from the unit sphere in $\mathbb{R}^d$, and then setting

$$\psi_k = \alpha \cdot \phi_k + \omega \cdot \frac{Q(\phi_k) z_k}{\|Q(\phi_k) z_k\|_2}$$

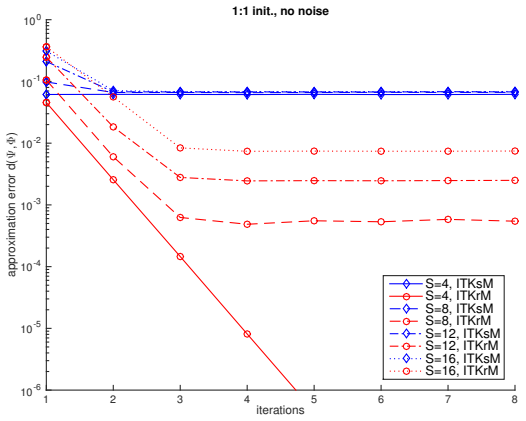
for the ratios $\alpha : \omega = 1 : 1$ and $\alpha : \omega = 1 : 4$. We also consider the completely random initialisation $\psi_k = z_k$.

For each initialisation dictionary we then run 100 iterations of ITKsM and ITKrM with the true sparsity level and dictionary size as input parameters, each time using a new batch of 100000 noiseless, respectively noisy signals. Figure 1 shows the average convergence respectively recovery rates over 20 trials for the three types of initialisations, using noiseless or noisy signals and for sparsity levels $S = 4, 8, 12, 16$.

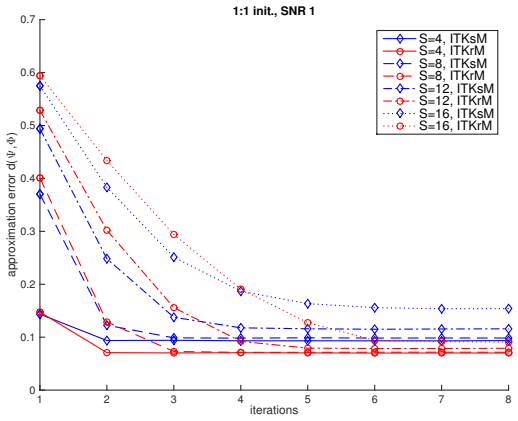
For the $1 : 1$ initialisations, despite the fact that the corresponding distance between the initialisations and the generating dictionary is much larger than the our estimated convergence radius, $d(\Psi, \Phi) = \sqrt{2 - \sqrt{2}} \approx 0.7654 \gg 1/\sqrt{\log K}$, both algorithms always converge to the generating dictionary, so we plot the distance $d(\Psi^{(n)}, \Phi)$ between the generating dictionary Φ and the output dictionary of the n-th iteration $\Psi^{(n)}$, Figure 1(a/b). As predicted by our theoretical results, using the same number of signals, ITKrM always leads to a more accurate estimate than ITKsM. As shown in Figure 1(a) for the noiseless signals with dynamic range 1 this difference is quite pronounced and especially in the case $S = 4 \leq \mu^{-1}/2$, the regime of unique sparsity, the precision of ITKrM is limited rather by the machine precision rather then the amount of training signals. From Figure 1(b) we see that both algorithms are locally stable even for the comparatively low signal to noise ratio $SNR = \mathbb{E}(\|\Phi x\|_2^2)/\mathbb{E}(\|r\|_2^2) = 1$ and coefficients with dynamic ranges varying between 1 and 0.9^{1-S}.

For the $1 : 4$ initialisations, corresponding to distance $d(\Psi, \Phi) = \sqrt{2 - 2/\sqrt{17}} \approx 1.2308$ between the initialisations and the generating dictionary, we do not always have convergence to the generating dictionary. We therefore plot the percentage of atoms recovered by the algorithm, using the convention that an atom ϕ_k is recovered if $\max_\ell |\langle \psi_\ell^{(n)}, \phi_k \rangle| \geq 0.99$, compare [3]. Counterintuitively to our theoretical results ITKrM turns out to be much more stable to far away initialisations. As we can see from Figure 1(c), in the case of noiseless signals ITKrM always recovers more than 99% of the atoms, while the recovery rate of ITKsM deteriorates quite drastically as the sparsity parameter S increases. To be more precise after 100 iterations ITKrM recovers the full dictionary for $17, 17, 15$ and 8 out of 20 initialisations for S taking values $4, 8, 12$ and 16 respectively, while ITKsM can only recover the full dictionary in case $S = 4$, (15 out of 20 initialisations), and for all other sparsity levels fails every time. The better performance of ITKrM is further confirmed by the results for noisy signals shown in Figure 1(d). While the recovery rates of ITKsM deteriorate further and even for $S = 4$ ITKsM can never recover the full dictionary, ITKrM continues to perform well. Indeed for ITKrM we can report a dithering effect, that is a better

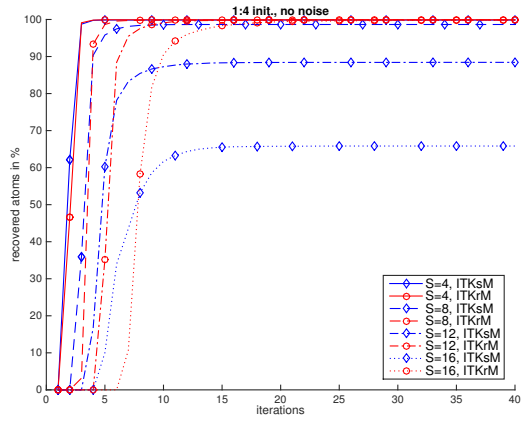
1. A Matlab Swiss knife (mini-toolbox) for playing with ITKrM and reproducing the experiments can be found at http://homepage.uibk.ac.at/~c7021041/code/ITKrM.zip.



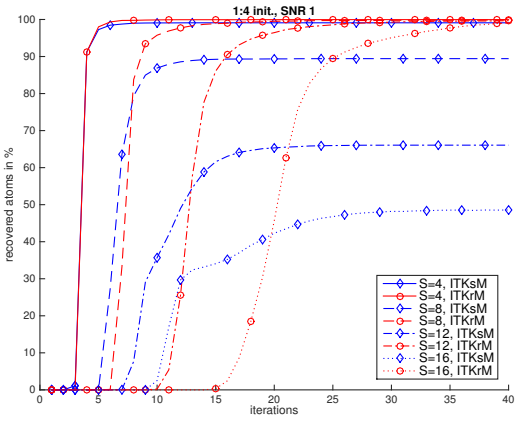
(a)



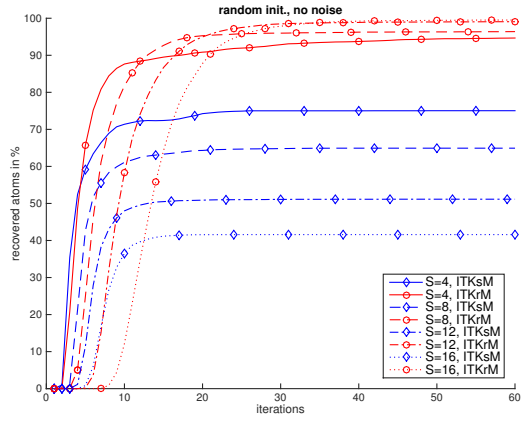
(b)



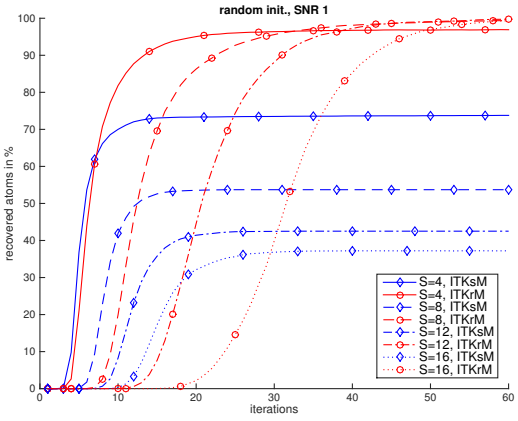
(c)



(d)



(e)



(f)

Fig. 1. Convergence respectively recovery rates of ITKsM and ITKrM for three initialisation types, corresponding to increasing distance to the generating dictionary, and using training signals with varying sparsity levels both in the noiseless and noisy case.

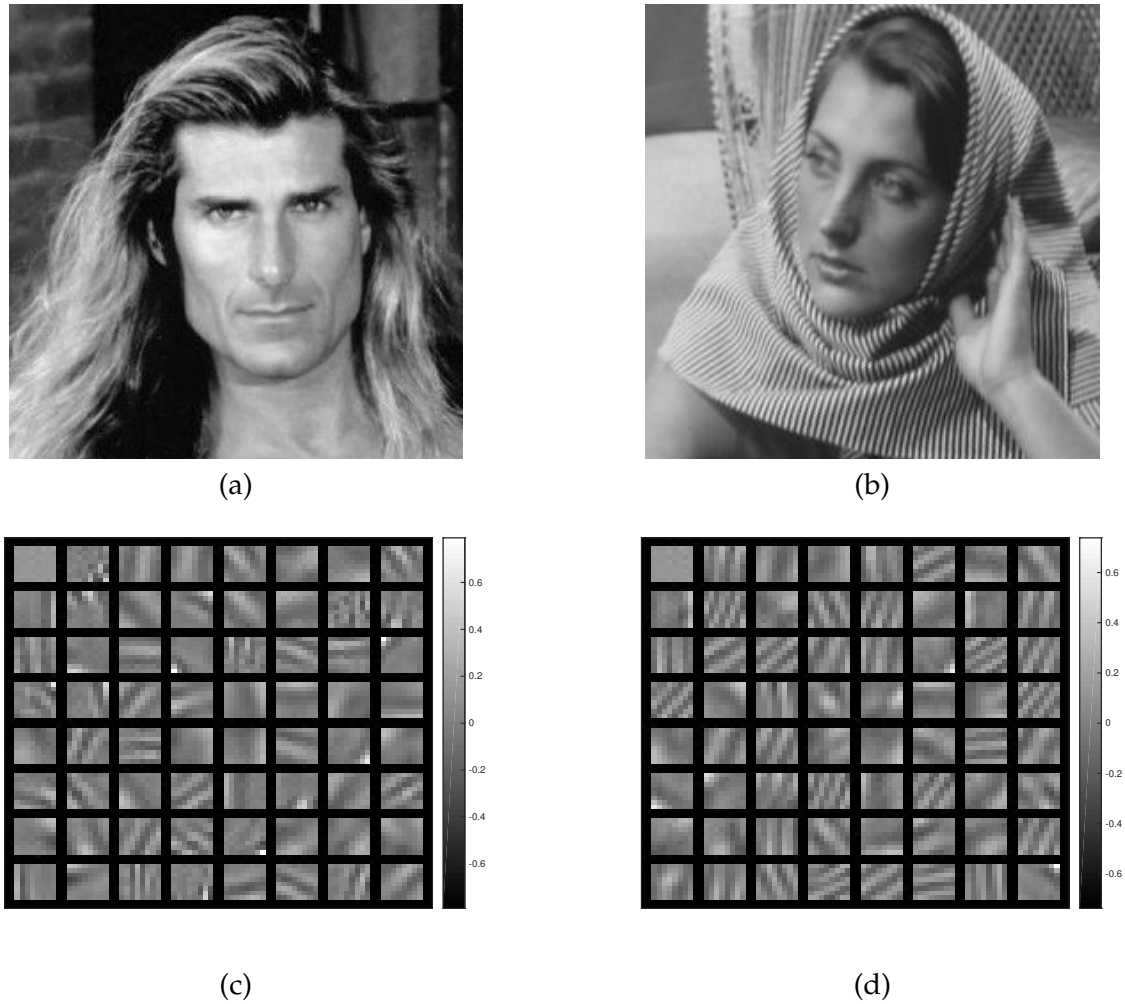
performance in noisy conditions, as ITKrM recovers the full dictionary 18 out of 20 times for $S = 4$ and always recovers the full dictionary for the other sparsity levels.
For the random initialisations we again plot the recovery rates, which confirm the trends observed for the $1 : 4$ initialisations, Figure 1(e/f). While the recovery rates of ITKsM are at best around 73% in the noiseless case for $S = 4$ decreasing to around 35% in the noisy case for $S = 16$, ITKrM always manages to recover at least 93% of the atoms. Interestingly even though again recovery speed decreases as S increases, both in the case of noiseless and noisy signals the recovery rates increase with S. So in the case of noiseless signals for $S = 12$ and $S = 16$ we again get a more than 99% recovery rate and can even report one respectively two full recoveries. In the case of noisy signals the dithering effect is now clearly visible and remarkably in case $S = 8$ ITKrM can recover the full dictionary 17 out of 20 times and for $S = 12, 16$ we always get full recovery.
Finally we also conduct a small experiment on image data to show that the more promising ITKrM algorithm is not merely a pretty toy for synthetic set-ups but indeed useful in practice. In particular for two 256×256 images, Fabio and Barbara, we take all 62001 possible 8×8 patches, normalise them and afterwards subtract their mean, that is we project the patches onto the orthogonal complement of the constant atom $\phi_1 \equiv 1/8$. On these patches we then learn a dictionary of 63 atoms, corresponding to the dimensionality of the signals after subtracting the mean (d=K=63). To be precise, we use a random initialisation, set the sparsity level $S = 5$, and in each of the 100 iterations use 10000 randomly selected patches. Figure 2 shows the two images together with their respective learned dictionaries (including the constant atom ϕ_1).

As we can see ITKrM is able to calculate meaningful dictionaries also on real data. In particular observe that even though we have used the same initialisation the dictionary learned on Barbara contains a lot more high frequency wave-like atoms, which capture the texture of the scarf. For the sake of conciseness we do not go into more details about the approximation performance of the learned dictionaries or possible image processing applications here but refer the interested reader to [23], [41], [36]. Instead we now turn to a final discussion of our results.

6 Discussion

We have shown that iterative thresholding and K-means is a very attractive dictionary learning method, since it has low computational complexity, $O(dKN)$ omitting log factors, can be used in parallel or online, has convergence radius $O(1/\sqrt{\log K})$ and sample complexity $O(K \log K \tilde{\varepsilon}^{-2})$ for a target error $\tilde{\varepsilon}$, which reduces to $O(K \log K \tilde{\varepsilon}^{-1})$ in the case of noiseless exactly sparse signals. Further to the best of our knowledge it is the only algorithm for learning overcomplete dictionaries, that is proven to be (locally) stable for sparsity ranges up to a log factor of the ambient dimension - that is recovery down to a target error $K^{-\ell}$ for sparsity levels S up to $O(\mu^{-2}/(\ell \log K)) = O(d/(\ell \log K)$.
As such it improves on related results in terms of computational efficiency, convergence radius and admissible sparsity level, [1], or in terms of achievable error and admissible sparsity level, [5]. In the case of noiseless signals, which is the only valid regime for [1], the sample complexity is in comparison larger by a factor ε^{-1}. However, note that there are information theoretic results indicating that in the case of noisy signals the dependence of the sample complexity on the inverse squared target error ε^{-2} is optimal, [25]. For an overview of results for iterative dictionary learning algorithms see Table 1. For a more general overview of theoretical results in dictionary learning see Table 1 in [19].

Further we have shown that in synthetic experiments the computationally more involved algorithm ITKrM often converges globally, when initalized with a random dictionary. This together with the fact that the algorithm is also able to learn meaningful dictionaries on image data makes it an attractive low complexity alternative to K-SVD.
The global convergence behaviour of ITKrM comes partly as a surprise since for ITKrM we can only prove a convergence radius of the order $O(1/\sqrt{S})$ as opposed to $O(1/\sqrt{\log K})$ for ITKsM. It also indicates that one might be able to increase the convergence radius of the algorithms by making additional assumptions on the perturbation dictionary, that is the normalised difference between the input and the generating dictionary, such as good conditioning and incoherence like the random perturbations in our experiments.



(a) (b)

(c) (d)

Fig. 2. 256×256 images together with the dictionaries learned on their 8×8 patches, (a,c) Fabio, (b,d) Barbara.

All that then remains to show is that most perturbations have this additional property and that one iteration of ITKrM conserves the property respectively that an additional small corrective step can restore the property.

For the theoretical results, another slight disappointment hidden in the O notation is, that both the convergence radius and implicitly also the limiting precision decrease with the dynamic range of the coefficients. This seems unavoidable since the success of thresholding depends on the dynamic range. So while we could improve our results to depend on an average dynamic range instead of the worst dynamic range by assuming a probability distribution on the dynamic range in our proofs, this average dynamic range will remain a limitation. To remove the dependence on the dynamic range we would have to replace thresholding by another sparse approximation method such as (Orthogonal) Matching Pursuit or Basis Pursuit, which is used in [1]. However, the only method that is known to be on average stable for sparsity levels $S \geq \sqrt{d}$ is Basis Pursuit, [46], and it will need some work to extend the corresponding results to perturbed dictionaries, noise and approximation error all at the same time. A maybe less daunting strategy is to extend the stability results for thresholding to iterative (hard) thresholding methods, [9], [10], [16], [24]. Another strategy to overcome large dynamic ranges, we are interested in, which would at the same time remove the requirement of knowing the exact sparsity level, is to extend our results to the case where we can only assure a gap between c_S and c_{S+T} for $T > 1$.

The most important research directions are concerned with the globality of the results. To get to an efficient algorithm we need to find initialisation strategies, such as in [5], that remain cost efficient also for sparsity levels $S = O(\mu^{-2}/(\ell \log K))$. An alternative strategy, we are currently pursuing, is based on

	online parallelisable	noise stability	convergence radius	admissible sparsity S	sample complexity	achievable error $\varepsilon \nearrow$.
Agarwal et.al. [1]	✗	✗	S^{-2}	$\min\{\mu^{-1}, d^{1/6}\}$	(K^2/S)	(0)
Arora et.al. [5]	✓	✓	$(\log d)^{-1}$	μ^{-1}	SK^*	$\sqrt{S/d}^*$
ITKsM	✓	✓	$(\log K)^{-1/2}$	μ^{-2}	$K\tilde{\varepsilon}^{-2}$	$K^{-\ell}\,(0)$
ITKrM	✓	✓	$S^{-1/2}$	μ^{-2}	$K\tilde{\varepsilon}^{-2}(K\tilde{\varepsilon}^{-1})$	$K^{-\ell}\,(0)$

To be read as $O(\cdot)$, non-leading log-factors omitted, noiseless case with $S \leq \mu^{-1}$ in brackets.
*valid for Algorithm 2. Algorithm 5 seems to achieve similar errors as ITKsM/ITKrM but at significantly higher computational cost. Further the dependence of its sample complexity on the target error is not made explicit.

TABLE 1
Comparison of theoretical results for iterative dictionary learning algorithms.

the earlier mentioned additional assumptions. If the perturbation dictionary not only has a flat spectrum but is itself incoherent and incoherent to the generating dictionary we expect one step of ITKrM to reduce the perturbation sizes but to keep the perturbation directions roughly the same. Estimating the volume of 'good' perturbations we could then calculate the probability that a random initialisation is successful or, in case this probability is too small, add a corrective step that restores the good properties of the current iterate.

ACKNOWLEDGEMENTS

This work was supported by the Austrian Science Fund (FWF) under Grant no. J3335 and Grant no. Y760 In addition the numerical simulations were supported by the Austrian Ministry of Science (BMWF) as part of the UniInfrastrukturprogramm of the Focal Point Scientific Computing at the University of Innsbruck. Thanks go also to the reviewers for their corrections and helpful suggestions and to the Computer Vision Laboratory of the University of Sassari, Italy, which provided the beautiful surroundings, where the inspirational part of the presented work was done.

APPENDIX A
PROOF SKETCHES

A.1 Proof of Corollary 3.3

The proof is analogue to the one of Theorem 3.2. We only need to take into account that without noise we have $C_r = 1$ and that in all estimates the constant $B+1$ can be replaced by B, since for noise free signals $y_n = \Phi x_{c_n,p_n,\sigma_n}$ we have $\|y_n\|_2 \leq B$. Further since the coefficients are strongly S-sparse, thresholding using the generating dictionary Φ will always (almost surely) recover the generating support with a margin $u_s \geq (\Delta_S - 2\mu S)c_n(1)$, that is $\min_{k\in I_n} |\langle \phi_k, y_n\rangle| \geq \max_{k\notin I_n} |\langle \phi_k, y_n\rangle| + u_s$, compare [40]. Therefore the event that thresholding using Ψ fails or that the empirical signs differ from the generating ones is contained in

$$\mathcal{F}_n^s := \Big\{ y_n : \exists k \text{ s.t. } \omega_k \Big| \sum_j \sigma_n(j) c_n\big(p_n(j)\big) \langle \phi_j, z_k\rangle \Big| \geq \frac{u_s - \frac{\varepsilon^2 c_n(S)}{2}}{2} \Big\} \tag{52}$$

and we get

$$\Big\| \bar{\psi}_k - \frac{\gamma_{1,S}}{K}\phi_k \Big\|_2 \leq \frac{2\sqrt{B}}{N} \sharp\{n : y_n \in \mathcal{F}_n^s\} + \left\| \frac{1}{N} \sum_n y_n \, \sigma_n(k) \, \chi(I_n, k) - \frac{\gamma_{1,S}}{K}\phi_k \right\|_2, \tag{53}$$

which can be estimated as before.

A.2 Proof of Theorem 4.2

As already mentioned we use the same two step procedure and ideas as in the proof of Theorem (3.2). *Step 1:* We first check how often thresholding with Ψ fails. Assuming thresholding recovers the generating support we show that the difference of the residuals using Φ or Ψ concentrates around its expectation, which is small. Finally we show that the sum of residuals using Φ converges to a scaled version of ϕ_k. To make the ideas precise we define the thresholding residual based on Ψ

$$R^t(\Psi, y_n, k) := \big[y_n - P(\Psi_{I^t_{\Psi,n}}) y_n + P(\psi_k) y_n \big] \cdot \operatorname{sign}(\langle \psi_k, y_n\rangle) \cdot \chi(I^t_{\Psi,n}, k) \tag{54}$$

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

$$R^o(\Psi, y_n, k) := \big[y_n - P(\Psi_{I_n}) y_n + P(\psi_k) y_n \big] \cdot \sigma_n(k) \cdot \chi(I_n, k). \tag{55}$$

We can now write,

$$\begin{aligned}
\bar{\psi}_k &= \frac{1}{N}\sum_n \big[R^t(\Psi, y_n, k) - R^o(\Psi, y_n, k)\big] + \frac{1}{N}\sum_n \big[R^o(\Psi, y_n, k) - R^o(\Phi, y_n, k)\big] + \frac{1}{N}\sum_n R^o(\Phi, y_n, k) \\
&= \frac{1}{N}\sum_n \big[R^t(\Psi, y_n, k) - R^o(\Psi, y_n, k)\big] + \frac{1}{N}\sum_n \big[R^o(\Psi, y_n, k) - R^o(\Phi, y_n, k)\big] \\
&\qquad + \frac{1}{N}\sum_n \big[y_n - P(\Phi_{I_n}) y_n \big] \cdot \sigma_n(k) \cdot \chi(I_n, k) + \left(\frac{1}{N}\sum_n \langle y_n, \phi_k\rangle \cdot \sigma_n(k) \cdot \chi(I_n, k) \right) \phi_k.
\end{aligned} \tag{56}$$

Abbreviating $s_k = \frac{1}{N}\sum_n \langle y_n, \phi_k\rangle \cdot \sigma_n(k) \cdot \chi(I_n, k)$ we get

$$\begin{aligned}
\|\bar{\psi}_k - s_k\phi_k\|_2 \leq \frac{1}{N} \Big\| \sum_n \big[R^t(\Psi, y_n, k) - R^o(\Psi, y_n, k)\big] \Big\|_2 \\
+ \frac{1}{N} \Big\| \sum_n \big[R^o(\Psi, y_n, k) - R^o(\Phi, y_n, k)\big] \Big\|_2 \\
+ \frac{1}{N} \Big\| \sum_n \big[y_n - P(\Phi_{I_n}) y_n \big] \cdot \sigma_n(k) \cdot \chi(I_n, k) \Big\|_2.
\end{aligned} \tag{57}$$

We first estimate the norm of the first sum using the fact that the operator $\mathbb{I}_d - P(\Psi_{I_n}) + P(\psi_k)$ is an orthogonal projection and that $\|y_n\|_2 \leq \sqrt{B+1}$,

$$\frac{1}{N}\Big\| \sum_n \left[R^t(\Psi, y_n, k) - R^o(\Psi, y_n, k)\right] \Big\|_2 \leq \frac{2\sqrt{B+1}}{N} \cdot \sharp\{n : R^t(\Psi, y_n, k) \neq R^o(\Psi, y_n, k)\}. \tag{58}$$

Next note that on the draw of y_n the event that the thresholding residual using Ψ is different from the oracle residual using Ψ, $\{y_n : R^t(\Psi, y_n, k) \neq R^o(\Psi, y_n, k)\}$ for any k is again contained in the events $\mathcal{E}_n \cup \mathcal{F}_n$ as defined in (27)/(28),

$$\{y_n : R^t(\Psi, y_n, k) \neq R^o(\Psi, y_n, k)\} \subseteq \{y_n : I^t_{\Psi,n} \neq I_n\} \cup \{y_n : \operatorname{sign}(\Psi^\star_{I_n} y_n) \neq \sigma_n(I_n)\} \subseteq \mathcal{E}_n \cup \mathcal{F}_n. \tag{59}$$

Substituting the corresponding bounds into (57) we get,

$$\begin{aligned}\|\bar{\psi}_k - s_k\phi_k\|_2 \leq{}& \frac{2\sqrt{B+1}}{N} \cdot \sharp\{n : y_n \in \mathcal{E}_n\} + \frac{2\sqrt{B+1}}{N} \cdot \sharp\{n : y_n \in \mathcal{F}_n\} \\ &+ \frac{1}{N}\Big\| \sum_n \left[R^o(\Psi, y_n, k) - R^o(\Phi, y_n, k)\right] \Big\|_2 + \frac{1}{N}\Big\| \sum_n \left[y_n - P(\Phi_{I_n})y_n\right] \cdot \sigma_n(k) \cdot \chi(I_n, k)\Big\|_2. \end{aligned} \tag{60}$$

For the first two terms on the right hand side we use the same estimates as in the proof of Theorem 3.2. To estimate the remaining two terms on the right hand side as well as s_k we use the corresponding lemmata in the appendix. From Lemma B.6 we know that

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_n \chi(I_n, k)\sigma_n(k)\langle y_n, \phi_k\rangle\right| \leq (1-t_0)\frac{C_r\gamma_{1,S}}{K}\right) \leq \exp\left(-\frac{Nt_0^2 C_r^2\gamma_{1,S}^2}{2K(1 + \frac{SB}{K} + S\rho^2 + t_0 C_r\gamma_{1,S}\sqrt{B+1}/3)}\right). \tag{61}$$

From Lemma B.8 we get that if $S \leq \min\{\frac{K}{98B}, \frac{1}{98\rho^2}\}$, $\varepsilon \leq \frac{1}{32\sqrt{S}}$ and $\varepsilon_\delta \leq \frac{1}{24(B+1)}$ then

$$\begin{aligned}&\mathbb{P}\left(\frac{1}{N}\left\|\sum_n [R^o(\Psi, y_n, k) - R^o(\Phi, y_n, k)]\right\|_2 \geq \frac{C_r\gamma_{1,S}}{K}(0.381\varepsilon + t_3)\right) \\ &\qquad \leq \exp\left(-\frac{t_3 C_r^2\gamma_{1,S}^2 N}{40K\max\{S, B+1\}} \min\left\{\frac{t_3}{\varepsilon^2 + \varepsilon_\delta\left(1 - \gamma_{2,S} + d\rho^2\right)/160}, \frac{5}{3}\right\} + \frac{1}{4}\right). \end{aligned} \tag{62}$$

Finally from Lemma B.7 we know that for $0 \leq t_4 \leq 1 - \gamma_{2,S} + d\rho^2$, we have

$$\begin{aligned}&\mathbb{P}\left(\left\|\frac{1}{N}\sum_n \left[y_n - P(\Phi_{I_n})y_n\right] \cdot \sigma_n(k) \cdot \chi(I_n, k)\right\|_2 \geq \frac{C_r\gamma_{1,S}}{K}\, t_4\right) \\ &\qquad \leq \exp\left(-\frac{t_4^2 C_r^2\gamma_{1,S}^2 N}{8K\max\{S, B+1\}\left(1 - \gamma_{2,S} + d\rho^2\right)} + \frac{1}{4}\right). \end{aligned} \tag{63}$$

Thus with high probability we have

$$\left\|\bar{\psi}_k - s_k\phi_k\right\|_2 \leq \frac{C_r\gamma_{1,S}}{K}\left(\varepsilon_{\mu,\rho} + t_1 + \tau\varepsilon + t_2 + 0.381\varepsilon + t_3 + t_4\right) \quad \text{and} \quad s_k \geq (1-t_0)\frac{C_r\gamma_{1,S}}{K}. \tag{64}$$

To be more precise, if we choose a target precision $\tilde{\varepsilon} \geq 8\varepsilon_{\mu,\rho}$ and set $t_1 = \tilde{\varepsilon}/24$, $t_2 = t_3 = \max\{\tilde{\varepsilon}, \varepsilon\}/24$, $\tau = 1/24$, $t_4 = \tilde{\varepsilon}/8$ and $t_0 = 1/50$ we get

$$\max_k \left\|\bar{\psi}_k - \frac{C_r\gamma_{1,S}}{K}\phi_k\right\|_2 \leq 0.8 \cdot \frac{C_r\gamma_{1,S}}{K}\max\{\tilde{\varepsilon}, \varepsilon\} \quad \text{and} \quad \min_k s_k \geq 0.98 \cdot \frac{C_r\gamma_{1,S}}{K}. \tag{65}$$

except with probability

$$\begin{aligned}&\exp\left(\frac{-C_r\gamma_{1,S}N\tilde{\varepsilon}}{336\,K\sqrt{B+1}}\right) + \exp\left(\frac{-C_r\gamma_{1,S}N\max\{\tilde{\varepsilon}, \varepsilon\}}{144\,K\sqrt{B+1}}\right) + K\exp\left(\frac{-C_r^2\gamma_{1,S}^2 N}{K(5103 + 34\,C_r\gamma_{1,S}\sqrt{B+1})}\right) \\ &+2K\exp\left(\frac{-C_r^2\gamma_{1,S}^2 N\tilde{\varepsilon}^2}{512K\max\{S, B+1\}\left(1 - \gamma_{2,S} + d\rho^2\right)}\right) + 2K\exp\left(\frac{-C_r^2\gamma_{1,S}^2 N\max\{\tilde{\varepsilon}, \varepsilon\}^2}{576K\max\{S, B+1\}\left(\varepsilon + 1 - \gamma_{2,S} + d\rho^2\right)}\right). \end{aligned}$$

Note that in case the target precision $\tilde{\varepsilon}$ is larger than ε_δ, as happens for instance as soon as $\beta_S \leq \frac{1}{7\sqrt{S}}$ and therefore $\varepsilon_{\mu,\rho} \geq \varepsilon_\delta$, the last term in the sum above reduces to

$$2K \exp\left(\frac{-C_r^2 \gamma_{1,S}^2 N \max\{\tilde{\varepsilon}, \varepsilon\}}{576K \max\{S, B+1\}\,(2 - \gamma_{2,S} + d\rho^2)}\right). \tag{66}$$

Lemma B.10 then again implies that

$$d(\bar{\Psi}, \Phi) = \max_k \left\| \frac{\bar{\psi}_k}{\|\bar{\psi}_k\|_2} - \phi_k \right\|_2 \leq 0.92 \max\{\tilde{\varepsilon}, \varepsilon\}. \tag{67}$$

Step 2: The second step is analogue to the one in the proof of Theorem 3.2.

A.3 Proof of Theorem 4.3

We follow the proof of Theorem 4.2 but take into account that in case of exactly S-sparse, noiseless signals the bound (57) reduces to

$$\|\bar{\psi}_k - s_k \phi_k\|_2 \leq \frac{2\sqrt{B}}{N} \cdot \sharp\{n : y_n \in \mathcal{F}_n^s\} + \frac{1}{N} \Big\| \sum_n \left[R^o(\Psi, y_n, k) - R^o(\Phi, y_n, k)\right] \Big\|_2. \tag{68}$$

Since the relative gap $\Delta_S > 2\mu S$ we get $\delta_S \leq \mu S \leq \frac{1}{2}$ and by Lemma B.4

$$\mathbb{P}\left(\sharp\{n : y_n \in \mathcal{F}_n^s\} \geq \frac{\gamma_{1,S} N}{2K\sqrt{B}} \cdot (\tau\varepsilon + t_2)\right) \leq \exp\left(\frac{-t_2^2 \gamma_{1,S} N}{2K\sqrt{B}\,(2\tau\varepsilon + t_2)}\right), \tag{69}$$

whenever

$$\varepsilon \leq \frac{\Delta_S - 2\mu S}{\sqrt{12}\left(\frac{1}{4} + \sqrt{\log\left(\frac{23K^2\sqrt{B}}{(\Delta_S - 2\mu S)\gamma_{1,S}\tau}\right)}\right)}. \tag{70}$$

Further by Lemma B.8

$$\mathbb{P}\left(\frac{1}{N}\left\|\sum_n \left[R^o(\Psi, y_n, k) - R^o(\Phi, y_n, k)\right]\right\|_2 \geq \frac{\gamma_{1,S}}{K}(1\varepsilon + t_3)\right) \leq \exp\left(-\frac{t_3 \gamma_{1,S}^2 N}{32\varepsilon K \max\{S, B\}} \min\left\{\frac{t_3}{\varepsilon}, 1\right\} + \frac{1}{4}\right),$$

and again by B.6

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_n \chi(I_n, k)\sigma_n(k)\langle y_n, \phi_k\rangle\right| \leq (1 - t_0)\frac{C_r \gamma_{1,S}}{K}\right) \leq \exp\left(-\frac{N t_0^2 \gamma_{1,S}^2}{2K(1 + \mu^2(S-1) + t_0\gamma_{1,S}\sqrt{B}/3)}\right). \tag{71}$$

Thus with high probability we have

$$\left\|\bar{\psi}_k - s_k\phi_k\right\|_2 \leq \frac{\gamma_{1,S}}{K}\left(\tau\varepsilon + t_2 + 0.611\varepsilon + t_3\right) \quad \text{and} \quad s_k \geq (1 - t_0)\frac{\gamma_{1,S}}{K}. \tag{72}$$

The final result follows as before from setting $t_0 = 1/50$, $\tau = 1/24$, $t_2 = \max\{\tilde{\varepsilon}, \varepsilon\}/24$ and $t_3 = 2t_2$.

APPENDIX B
PROBABILITY ESTIMATES & TECHNICALITIES

Theorem B.1 (Vector Bernstein, [28], [22], [29]). *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 $0 \leq t \leq m_2/(M + m_1)$*

$$\mathbb{P}\left(\left\|\sum_n v_n - \sum_n \mathbb{E}(v_n)\right\|_2 \geq t\right) \leq \exp\left(-\frac{t^2}{8m_2} + \frac{1}{4}\right), \tag{73}$$

and in general

$$\mathbb{P}\left(\left\|\sum_n v_n - \sum_n \mathbb{E}(v_n)\right\|_2 \geq t\right) \leq \exp\left(-\frac{t}{8}\cdot\min\left\{\frac{t}{m_2}, \frac{1}{M+m_1}\right\} + \frac{1}{4}\right). \tag{74}$$

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 $m_2 = t(M+m_1)$.
For the simple case of random variables we also state a scalar version of Bernstein's inequality leading to better constants.

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

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

Lemma B.3. *For y_n following model* (13) *with coefficients that have an absolute gap β_S we have,*

$$\mathbb{P}\left(\sharp\{n : y_n \in \mathcal{E}_n\} \geq \frac{C_r\gamma_{1,S}N}{2K\sqrt{B+1}}\cdot(\varepsilon_{\mu,\rho} + t)\right) \leq \exp\left(\frac{-t^2 C_r\gamma_{1,S}N}{2K\sqrt{B+1}\,(2\varepsilon_{\mu,\rho}+t)}\right), \tag{75}$$

where $\varepsilon_{\mu,\rho} = \frac{8K^2\sqrt{B+1}}{C_r\gamma_{1,S}}\exp\left(\frac{-\beta_S^2}{98\max\{\mu^2,\rho^2\}}\right)$.

Proof: We apply Theorem B.2 to the sum of recentered indicator functions $\mathbf{1}_{\mathcal{E}_\mathbf{n}} - \mathbb{P}(\mathcal{E}_\mathbf{n})$ to get

$$\mathbb{P}\left(\sharp\{n : y_n \in \mathcal{E}_n\} \geq \sum_n \mathbb{P}(\mathcal{E}_n) + tN\right) \leq \exp\left(\frac{-t^2N^2}{2\sum_n \mathbb{P}(\mathcal{E}_n) + tN}\right). \tag{76}$$

To estimate $\mathbb{P}(\mathcal{E}_n)$ we apply Hoeffding's inequality to (27) resp. use the subgaussian property of r_n. Omitting subscripts for simplicity and abbreviating $u = c(S) - c(S+1)$ we get,

$$\begin{aligned}
\mathbb{P}(\mathcal{E}) &\leq \sum_k \mathbb{P}\left(\Big|\sum_{j\neq k}\sigma(j)c\big(p(j)\big)\langle\phi_j,\phi_k\rangle\Big| \geq \frac{u}{7}\right) + \sum_k \mathbb{P}\left(|\langle r,\phi_k\rangle| \geq \frac{u}{7}\right)\\
&\leq \sum_k 2\exp\left(\frac{u^2}{98\sum_{j\neq k} c\big(p(j)\big)^2|\langle\phi_j,\phi_k\rangle|^2}\right) + 2K\exp\left(\frac{-u^2}{98\rho^2}\right)\\
&\leq 2K\exp\left(\frac{-\beta_S^2}{98\mu^2}\right) + 2K\exp\left(\frac{-\beta_S^2}{98\rho^2}\right)\\
&\leq 4K\exp\left(\frac{-\beta_S^2}{98\max\{\mu^2,\rho^2\}}\right) = \frac{C_r\gamma_{1,S}}{2K\sqrt{B+1}}\cdot\varepsilon_{\mu,\rho}.
\end{aligned} \tag{77}$$

The result follows from the substitution $t \to \frac{C_r\gamma_{1,S}}{2K\sqrt{B+1}}\,t$.

□

Lemma B.4. *(a) For y_n following model* (13) *with coefficients that have a relative gap Δ_S we have,*

$$\mathbb{P}\left(\sharp\{n : y_n \in \mathcal{F}_n\} \geq \frac{C_r\gamma_{1,S}N}{2K\sqrt{B+1}}\cdot(\tau\varepsilon + t)\right) \leq \exp\left(\frac{-t^2 C_r\gamma_{1,S}N}{2K\sqrt{B+1}\,(2\tau\varepsilon+t)}\right), \tag{78}$$

whenever

$$\varepsilon \leq \frac{\Delta_S}{\sqrt{98B}\left(\frac{1}{4} + \sqrt{\log\left(\frac{106K^2(B+1)}{\Delta_S C_r\gamma_{1,S}\tau}\right)}\right)}. \tag{79}$$

(b) For y_n following model (13) *with coefficients that have a relative gap $\Delta_S \geq 2\mu S$ we have,*

$$\mathbb{P}\left(\sharp\{n : y_n \in \mathcal{F}_n^s\} \geq \frac{\gamma_{1,S} N}{2K\sqrt{B}} \cdot (\tau\varepsilon + t)\right) \leq \exp\left(\frac{-t^2 \gamma_{1,S} N}{2K\sqrt{B}\,(2\tau\varepsilon + t)}\right), \tag{80}$$

whenever

$$\varepsilon \leq \frac{\Delta_S - 2\mu S}{\sqrt{8B}\left(\frac{1}{4} + \sqrt{\log\left(\frac{19K^2 B}{(\Delta_S - 2\mu S)\gamma_{1,S}\tau}\right)}\right)}. \tag{81}$$

Proof: We apply Theorem B.2 to the sum of recentered indicator functions $\mathbf{1}_{\mathcal{F}_\mathbf{n}^{(\mathbf{s})}} - \mathbb{P}(\mathcal{F}_\mathbf{n}^{(\mathbf{s})})$ to get

$$\mathbb{P}\left(\sharp\{n : y_n \in \mathcal{F}_n^{(s)}\} \geq \sum_n \mathbb{P}(\mathcal{F}_n^{(s)}) + tN\right) \leq \exp\left(\frac{-t^2 N^2}{2\sum_n \mathbb{P}(\mathcal{F}_n^{(s)}) + tN}\right). \tag{82}$$

To estimate $\mathbb{P}(\mathcal{F}_n^{(s)})$ we again apply Hoeffding's inequality this time to (28)/(52) resp. use the subgaussian property of r_n. Omitting subscripts and using the short hand $u = c(S) - c(S+1)$ and $u_s = (\Delta_S - 2\mu S)c(1)$ we get,

$$\begin{aligned}
\mathbb{P}(\mathcal{F}) &\leq \sum_k \mathbb{P}\left(\omega_k \Big| \sum_{j\neq k} \sigma(j) c\big(p(j)\big)\langle \phi_j, z_k\rangle\Big| \geq \frac{u}{7} - \frac{\varepsilon^2 c(S)}{6}\right) + \sum_k \mathbb{P}\left(\omega_k |\langle r, z_k\rangle| \geq \frac{u}{14} - \frac{\varepsilon^2 c(S)}{12}\right) \\
&\leq \sum_k 2\exp\left(\frac{-\left(u - \frac{7\varepsilon^2 c(S)}{6}\right)^2}{98\omega_k^2 \sum_{j\neq k} c\big(p(j)\big)^2 |\langle \phi_j, z_k\rangle|^2}\right) + 2K\exp\left(\frac{-\left(u - \frac{7\varepsilon^2 c(S)}{6}\right)^2}{4\cdot 98\rho^2}\right) \\
&\leq 2K\exp\left(\frac{-\left(u - \frac{7\varepsilon^2 c(S)}{6}\right)^2}{98\varepsilon^2 \min\{c(1)^2 B, 1\}}\right) + 2K\exp\left(\frac{-\left(u - \frac{7\varepsilon^2 c(S)}{6}\right)^2}{4\cdot 98\varepsilon^2\rho^2}\right) \\
&\leq 5K\exp\left(\frac{-(c(S) - c(S+1))^2}{98\varepsilon^2 c(1)^2 B}\right) \leq 5K\exp\left(\frac{-\Delta_S^2}{98\varepsilon^2 B}\right).
\end{aligned} \tag{83}$$

From Lemma A.3 in [39] we further know that condition (79) implies

$$5K\exp\left(\frac{-\Delta_S^2}{98\varepsilon^2 B}\right) \leq \frac{C_r \gamma_{1,S}}{2K\sqrt{B+1}} \cdot \tau\varepsilon, \tag{84}$$

and the result in (a) follows again from the substitution $t \to \frac{C_r\gamma_{1,S}}{2K\sqrt{B+1}}\, t$.
Similarly we get

$$\begin{aligned}
\mathbb{P}(\mathcal{F}^s) &\leq \sum_k \mathbb{P}\left(\omega_k \Big| \sum_{j\neq k} \sigma(j) c\big(p(j)\big)\langle \phi_j, z_k\rangle\Big| \geq \frac{u_s}{2} - \frac{\varepsilon^2 c(S)}{4}\right) \\
&\leq 2K\exp\left(\frac{-\left((\Delta_S - 2\mu S)c(1) - \frac{\varepsilon^2 c(S)}{2}\right)^2}{8\varepsilon^2 \min\{c(1)^2 B, 1\}}\right) \leq 3K\exp\left(\frac{-(\Delta_S - 2\mu S)^2}{8\varepsilon^2 B}\right) \leq \frac{\gamma_{1,S}}{2K\sqrt{B}} \cdot \tau\varepsilon,
\end{aligned} \tag{85}$$

whenever (81) holds and the result in (b) follows from the substitution $t \to \frac{\gamma_{1,S}}{2K\sqrt{B}}\, t$.
Finally note that another (messier) way to bound $\sum_{j\neq k} c\big(p(j)\big)^2 |\langle \phi_j, z_k\rangle|^2$ is

$$\sum_{j\neq k} c\big(p(j)\big)^2 |\langle \phi_j, z_k\rangle|^2 \leq \min\{c(1)^2 \|\Phi_I\|_{2,2}^2 + 1 - \gamma_{2,S}, c(1)^2\|\Phi_I\|_{2,2}^2 + c(S+1)^2 B\}. \tag{86}$$

However, in the case of exactly S-sparse signals these can lead to better (and again clean) estimates, such as $c(1)^2(1+\mu S)$ or $c(1)^2(1+\delta_S)$ if Φ has isometry constant $\delta_S < 1$. □

Remark B.1. The last two lemmata are used to prove that, once the perturbed dictionary Ψ is within radius $O(1/\log(K))$ of the generating dictionary Φ, thresholding will always succeed in recovering the full generating support, even for $S = O(\mu^{-2})$. Without assuming random signs, we can still get that thresholding recovers the generating support once Ψ is within radius $O(1/\sqrt{S})$ for reduced sparsity levels $S = O(\mu^{-1})$.

Lemma B.5. *For* $y_n = \frac{\Phi x_{c_n,p_n,\sigma_n} + r_n}{\sqrt{1+\|r_n\|_2^2}}$ *as in model* (13) *and* $0 \leq t \leq \frac{\sqrt{S}}{\sqrt{B}+2}$ *we have*

$$\mathbb{P}\left(\left\|\frac{1}{N}\sum_n \frac{\Phi x_{c_n,p_n,\sigma_n} + r_n}{\sqrt{1+\|r_n\|_2^2}} \cdot \sigma_n(k) \cdot \chi(I_n,k) - \frac{C_r\gamma_{1,S}}{K}\phi_k\right\|_2 \geq \frac{C_r\gamma_{1,S}}{K}\, t\right) \leq \exp\left(-\frac{t^2 C_r^2 \gamma_{1,S}^2 N}{8SK} + \frac{1}{4}\right). \quad (87)$$

Proof: We apply Theorem B.1 to $v_n = \frac{\Phi x_{c_n,p_n,\sigma_n} + r_n}{\sqrt{1+\|r_n\|_2^2}} \cdot \sigma_n(k) \cdot \chi(I_n,k)$. Since the v_n are identically distributed we drop the index n for our estimates. Remembering that $I = p^{-1}(\mathbb{S})$ we get,

$$\begin{aligned}
\mathbb{E}(v) &= \mathbb{E}_{c,p,\sigma,r}\left(\frac{\chi(I,k)}{\sqrt{1+\|r\|_2^2}}\left(\sum_j \phi_j c\big(p(j)\big)\sigma(j)\cdot\sigma(k) + r\cdot\sigma(k)\right)\right) \\
&= \mathbb{E}_{c,p,r}\left(\frac{\chi(\mathbb{S},p(k))\cdot c\big(p(k)\big)}{\sqrt{1+\|r\|_2^2}}\,\phi_k\right) \\
&= \mathbb{E}_r\left(\frac{1}{\sqrt{1+\|r\|_2^2}}\right)\mathbb{E}_c\left(\frac{c(1)+\ldots+c(S)}{K}\right)\phi_k = \frac{C_r\gamma_{1,S}}{K}\,\phi_k, \qquad (88)
\end{aligned}$$

and $\|\mathbb{E}(v)\|_2 \leq \sqrt{S}/K$. Together with the estimates,

$$\begin{aligned}
&\mathbb{E}\left(\|v\|_2^2\right) = \mathbb{E}\left(\frac{\chi(I,k)}{1+\|r\|_2^2}\cdot\left(\|\Phi x_{c,p,\sigma}\|_2^2 + \langle \Phi x_{c,p,\sigma}, r\rangle + \|r\|_2^2\right)\right) = \mathbb{E}\left(\chi(I,k)\right) = \frac{S}{K} \\
&\text{and} \qquad \|v\|_2 \leq \frac{\|\Phi x_{c,p,\sigma} + r\|_2}{\sqrt{1+\|r\|_2^2}} \leq \frac{\sqrt{B}+\|r\|_2}{\sqrt{1+\|r\|_2^2}} \leq \sqrt{B+1},
\end{aligned}$$

this leads to

$$\mathbb{P}\left(\left\|\frac{1}{N}\sum_n \frac{\Phi x_{c_n,p_n,\sigma_n} + r_n}{\sqrt{1+\|r_n\|_2^2}} \cdot \sigma_n(k) \cdot \chi(I_n,k) - \frac{C_r\gamma_{1,S}}{K}\phi_k\right\|_2 \geq t\right) \leq \exp\left(-\frac{t^2 K N}{8S} + \frac{1}{4}\right), \quad (89)$$

for $0 \leq t \leq \frac{S}{K(\sqrt{B+1}+\frac{S}{K})}$. The final statements follows from the substitution $t \to \frac{C_r\gamma_{1,S}}{K}\, t$ and simplifications. □

Remark B.2. Note that for Eq. (88) in the above proof, we have used the sign invariance in our model but not the permutation invariance. For very small sparsity levels we can also get a stable version of the lemma using only the permutation invariance. Assume for simplicity that Φ is an orthonormal basis and that the sparse coefficients are constant, $c_k \equiv c$ for $k \leq S$ and zero else. In this worst case scenario where the signs never cancel out we get

$$\mathbb{E}(v) = c\left(\phi_k + \frac{S-1}{d-1}\sum_{j\neq k}\phi_j\right) \qquad \text{and} \qquad \|\mathbb{E}(v)\|_2 = c\sqrt{1+\frac{(S-1)^2}{d-1}}, \quad (90)$$

which implies that the atoms can be learned up to a precision $O(S^2/d)$. A relaxed condition replacing sign and permutation invariance could be that the coefficient sequences x satisfy $\mathbb{E}\left(x(j)\,\mathrm{sign}(x(k))|k \in I\right) \ll \mathbb{E}(|x(k)|\,|k \in I)$ for I containing the indices of the S largest coordinates in absolute value, that is $\min_{i\in I}|x(i)| > \max_{j\notin I}|x(j)|$. This condition is quite natural as it basically prevents two atoms ϕ_k and ϕ_j from always appearing together in the same ratio $x(k) : x(j) = a : b$. In this case they could simply be replaced by two copies of the same atom, $\tilde{\phi}_j = \tilde{\phi}_k = a\phi_k + b\phi_j$ which would increase the response criterion on which ITKsM is based, see [40].

Lemma B.6. *For* $y_n = \frac{\Phi x_{c_n,p_n,\sigma_n} + r_n}{\sqrt{1+\|r_n\|_2^2}}$ *as in model* (13) *we have*

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_n \chi(I_n,k)\sigma_n(k)\langle y_n,\phi_k\rangle\right| \le (1-t)\frac{C_r\gamma_{1,S}}{K}\right) \le \exp\left(-\frac{Nt^2C_r^2\gamma_{1,S}^2}{2K(1+\frac{SB}{K}+S\rho^2+tC_r\gamma_{1,S}\sqrt{B+1}/3)}\right). \tag{91}$$

Proof: We apply Theorem B.2 to $v_n - \mathbb{E}(v_n)$ for $v_n = \chi(I_n,k)\sigma_n(k)\langle y_n,\phi_k\rangle$, as usual dropping the index n in the estimates for conciseness. For the expectation we get

$$\begin{aligned}\mathbb{E}(v) &= \mathbb{E}_{c,p,\sigma,r}\left(\frac{\chi(I,k)}{\sqrt{1+\|r\|_2^2}}\left(\sum_j c\big(p(j)\big)\sigma(j)\langle\phi_j,\phi_k\rangle\cdot\sigma(k)+\langle r,\phi_k\rangle\cdot\sigma(k)\right)\right)\\ &= \mathbb{E}_{c,p,r}\left(\frac{\chi(\mathbb{S},p(k))\cdot c\big(p(k)\big)}{\sqrt{1+\|r\|_2^2}}\right) = \frac{C_r\gamma_{1,S}}{K}.\end{aligned} \tag{92}$$

We further estimate the second moment of v as

$$\begin{aligned}\mathbb{E}\left(v^2\right) &= \mathbb{E}_{c,p,\sigma,r}\left(\frac{\chi(I,k)}{1+\|r\|_2^2}\Big(\sum_j c\big(p(j)\big)\sigma(j)\langle\phi_j,\phi_k\rangle+\langle r,\phi_k\rangle\Big)^2\right)\\ &\le \mathbb{E}_{c,p}\left(\chi(I,k)\cdot\left(\sum_j c\big(p(j)\big)^2|\langle\phi_j,\phi_k\rangle|^2+\mathbb{E}_r\left(|\langle r,\phi_k\rangle|^2\right)\right)\right)\\ &\le \mathbb{E}_{c,p}\left(\chi(I,k)\cdot\left(\frac{\gamma_{2,S}}{S}+\frac{1-\frac{\gamma_{2,S}}{S}}{K-1}\sum_{j\in I,j\neq k}|\langle\phi_j,\phi_k\rangle|^2+\rho^2\right)\right) \le \frac{S}{K}\cdot\left(\frac{\gamma_{2,S}}{S}+\frac{B}{K}+\rho^2\right).\end{aligned} \tag{93}$$

Thus we can choose $m = \frac{1}{K}\left(\gamma_{2,S}+\frac{BS}{K}+S\rho^2-\frac{C_r^2\gamma_{1,S}^2}{K}\right)$. Note that in the case of exactly S-sparse signals, where $\gamma_{2,S}=1$ we could also use an alternative bound based on the fact that $\mathbb{E}\left(v^2\right) \le \frac{1}{K}(1+(S-1)\mu^2+S\rho^2)$. Since $|v| \le |\langle y,\phi_k\rangle| \le \|y\|_2 \le \sqrt{B+1}$ we can choose $M = \frac{1}{3}\left(\sqrt{B+1}+\frac{C_r\gamma_{1,S}}{K}\right)$. The final statement follows from the substitution $t \to \frac{C_r\gamma_{1,S}}{K}\,t$ and simplifications. □

Lemma B.7. *For* $y_n = \frac{\Phi x_{c_n,p_n,\sigma_n} + r_n}{\sqrt{1+\|r_n\|_2^2}}$ *as in model* (13)

$$\begin{aligned}\mathbb{P}\left(\left\|\frac{1}{N}\sum_n (y_n - P(\Phi_{I_n})y_n)\cdot\sigma_n(k)\cdot\chi(I_n,k)\right\|_2 \ge \frac{C_r\gamma_{1,S}}{K}\,t\right)\\ \le \exp\left(-\frac{tC_r^2\gamma_{1,S}^2N}{8K\max\{S,B+1\}}\min\left\{\frac{t}{1-\gamma_{2,S}+d\rho^2},1\right\}+\frac{1}{4}\right).\end{aligned} \tag{94}$$

Proof: We apply Theorem B.1 to $v_n = (y_n - P(\Phi_{I_n})y_n)\cdot\sigma_n(k)\cdot\chi(I_n,k)$. For brevity we again drop the index n in the estimates and define the orthogonal projection $Q(\Phi_I) = \mathbb{I}_d - P(\Phi_I)$. For the expectation we get

$$\begin{aligned}\mathbb{E}(v) &= \mathbb{E}_{c,p,\sigma,r}\left(\frac{\chi(I,k)}{\sqrt{1+\|r\|_2^2}}\,Q(\Phi_I)\left(\sum_j \phi_j c\big(p(j)\big)\sigma(j)\cdot\sigma(k)+r\cdot\sigma(k)\right)\right)\\ &= \mathbb{E}_{c,p,r}\left(\frac{\chi(I,k)}{\sqrt{1+\|r\|_2^2}}\,c\big(p(k)\big)Q(\Phi_I)\phi_k\right) = 0,\end{aligned} \tag{95}$$

and for the second moment

$$\begin{aligned}\mathbb{E}\left(\|v\|_2^2\right) &= \mathbb{E}_{c,p,\sigma,r}\left(\frac{\chi(I,k)}{1+\|r\|_2^2}\cdot\left(\|Q(\Phi_I)\Phi x_{c,p,\sigma}\|_2^2+\langle Q(\Phi_I)\Phi x_{c,p,\sigma},Q(\Phi_I)r\rangle+\|Q(\Phi_I)r\|_2^2\right)\right)\\ &\leq \mathbb{E}_{c,p}\left(\chi(I,k)\cdot\left(\sum_j c\big(p(j)\big)^2\|Q(\Phi_I)\phi_j\|_2^2+\mathbb{E}_r\left(\frac{\|Q(\Phi_I)r\|_2^2}{1+\|r\|_2^2}\right)\right)\right)\\ &\leq \mathbb{E}_{c,p}\left(\chi(I,k)\cdot\left(\sum_{j\notin I} c\big(p(j)\big)^2+\min\{1,(d-S)\rho^2\}\right)\right)\leq\frac{S}{K}\cdot\left(1-\gamma_{2,S}+d\rho^2\right).\end{aligned}\tag{96}$$

Since v is bounded,

$$\|v\|_2\leq\frac{\|Q(\Phi_I)(\Phi x_{c,p,\sigma}+r)\|_2}{\sqrt{1+\|r\|_2^2}}\leq\frac{\sqrt{B(1-\gamma_{2,S,\min})}+\|r\|_2}{\sqrt{1+\|r\|_2^2}}\leq\sqrt{B(1-\gamma_{2,S,\min})+1}\leq\sqrt{B+1},\tag{97}$$

we get for $t\to\frac{C_r\gamma_{1,S}}{K}\,t$

$$\begin{aligned}&\mathbb{P}\left(\left\|\frac{1}{N}\sum_n(y_n-P(\Phi_{I_n})y_n)\cdot\sigma_n(k)\cdot\chi(I_n,k)\right\|_2\geq\frac{C_r\gamma_{1,S}}{K}\,t\right)\\ &\qquad\leq\exp\left(-\frac{tC_r\gamma_{1,S}N}{8K}\min\left\{\frac{tC_r\gamma_{1,S}}{S(1-\gamma_{2,S}+d\rho^2)},\frac{1}{\sqrt{B+1}}\right\}+\frac{1}{4}\right)\\ &\qquad\leq\exp\left(-\frac{tC_r^2\gamma_{1,S}^2N}{8K}\min\left\{\frac{t}{S(1-\gamma_{2,S}+d\rho^2)},\frac{1}{C_r\gamma_{1,S}\sqrt{B+1}}\right\}+\frac{1}{4}\right).\end{aligned}\tag{98}$$

The final bound follows from the fact that $C_r<1$ and $\gamma_{1,S}\leq\sqrt{S}$.

□

Remark B.3. Note that for the above lemma neither the sign nor the permutation invariance are crucial. Without both assumptions we could still get a stable version of the lemma because we can bound $\mathbb{E}(v)$ by the residual energy $\|y_n-P(\Phi_{I_n}y_n\|_2$, which should be small if the signals are assumed to be S-sparse. To get perfect recoverability $\mathbb{E}(v)$ we could make the natural assumption that in expectation the residuals $a_n=y_n-P(\Phi_{I_n})y_n=Q(\Phi_{I_n})\Phi x_n$ are uncorrelated with the sign of the k-th coefficient $x_n(k)$ whenever $k\in I_n$, $\mathbb{E}\left(a_n\operatorname{sign}(x(k))\chi(I_n,k)\right)=0$. Indeed if this is not the case it means that the signals can be even better sparsely approximated if the atom ϕ_k is distorted towards this signed residual mean.

Lemma B.8. *Assume that* $y_n=\frac{\Phi x_{c_n,p_n,\sigma_n}+r_n}{\sqrt{1+\|r_n\|_2^2}}$ *follows the random model in* (13)*. Assume* $S\leq\min\{\frac{K}{98B},\frac{1}{98\rho^2}\}$ *and* $d(\Phi,\Psi)=\varepsilon\leq\frac{1}{32\sqrt{S}}$.

(a) If $\varepsilon_\delta:=K\exp\left(-\frac{1}{4741\mu^2S}\right)\leq\frac{1}{48(B+1)}$ *we have*

$$\begin{aligned}&\mathbb{P}\left(\frac{1}{N}\left\|\sum_n\left[R^o(\Psi,y_n,k)-R^o(\Phi,y_n,k)\right]\right\|_2\geq\frac{C_r\gamma_{1,S}}{K}(0.381\varepsilon+t)\right)\\ &\qquad\leq\exp\left(-\frac{tC_r\gamma_{1,S}N}{8K}\min\left\{\frac{tC_r\gamma_{1,S}}{S\left[5\varepsilon^2+\varepsilon_\delta\left(1-\gamma_{2,S}+d\rho^2\right)/32\right]},\frac{1}{3\sqrt{B+1}}\right\}+\frac{1}{4}\right).\end{aligned}\tag{99}$$

(b) If $\gamma_{2,S}=1,\rho=0$ *together with* $\varepsilon_\delta\leq\frac{1}{48(B+1)}$ *or* $\delta_S(\Phi)\leq1/4$ *this reduces to*

$$\begin{aligned}&\mathbb{P}\left(\frac{1}{N}\left\|\sum_n\left[R^o(\Psi,y_n,k)-R^o(\Phi,y_n,k)\right]\right\|_2\geq\frac{C_r\gamma_{1,S}}{K}(0.381\varepsilon+t)\right)\\ &\qquad\leq\exp\left(-\frac{t\gamma_{1,S}^2N}{32\varepsilon K\max\{S,B\}}\min\left\{\frac{t}{\varepsilon},1\right\}+\frac{1}{4}\right).\end{aligned}\tag{100}$$

(c) If $\gamma_{2,S} = 1, \rho = 0$ and $\delta_S(\Phi) \leq 1/2$ we have

$$\mathbb{P}\left(\frac{1}{N}\left\|\sum_n [R^o(\Psi, y_n, k) - R^o(\Phi, y_n, k)]\right\|_2 \geq \frac{\gamma_{1,S}}{K}(0.611\varepsilon + t)\right) \leq \exp\left(-\frac{t\gamma_{1,S}^2 N}{32\varepsilon K \max\{S,B\}} \min\left\{\frac{t}{\varepsilon}, 1\right\} + \frac{1}{4}\right). \tag{101}$$

Proof: We apply Theorem B.1 to $v_n = R^o(\Psi, y_n, k) - R^o(\Phi, y_n, k)$. Again we drop the index n in the estimates. Remembering the definition of $R^o(\Psi, y_n, k)$ in (55) we first expand v as

$$\begin{aligned} v &= \big(y_n - P(\Psi_{I_n})y_n + P(\psi_k)y_n\big) \cdot \sigma_n(k) \cdot \chi(I_n, k) - \big(y_n - P(\Phi_{I_n})y_n + P(\phi_k)y_n\big) \cdot \sigma_n(k) \cdot \chi(I_n, k) \\ &= \left[P(\Phi_I) - P(\Psi_I) - P(\phi_k) + P(\psi_k)\right] y \cdot \sigma(k) \cdot \chi(I, k). \end{aligned} \tag{102}$$

Abbreviate $T(I,k) := P(\Phi_I) - P(\Psi_I) - P(\phi_k) + P(\psi_k)$. Taking the expectation we get

$$\begin{aligned} \mathbb{E}(v) &= \mathbb{E}_{c,p,\sigma,r}\left(\frac{\chi(I,k)}{\sqrt{1+\|r\|_2^2}}\, T(I,k)\left(\sum_j \phi_j c\big(p(j)\big)\sigma(j) \cdot \sigma(k) + r \cdot \sigma(k)\right)\right) \\ &= \mathbb{E}_{c,p,r}\left(\frac{\chi(I,k) \cdot c\big(p(k)\big)}{\sqrt{1+\|r\|_2^2}}\left[P(\Phi_I) - P(\Psi_I) - P(\phi_k) + P(\psi_k)\right]\phi_k\right) \\ &= \frac{C_r\gamma_{1,S}}{K}\binom{K-1}{S-1}^{-1} \sum_{|I|=S, k\in I} \left[P(\psi_k) - P(\Psi_I)\right]\phi_k. \end{aligned} \tag{103}$$

We next split the sum above into a sum over the well-conditioned subsets, where $\delta_I(\Phi) \leq \delta_0$, and the ill-conditioned subsets, $\delta_I(\Phi) > \delta_0$,

$$\mathbb{E}(v) = \frac{C_r\gamma_{1,S}}{K}\binom{K-1}{S-1}^{-1}\left(\sum_{\substack{|I|=S,k\in I \\ \delta(\Phi_I)\leq\delta_0}} \left[P(\psi_k) - P(\Psi_I)\right]\phi_k + \sum_{\substack{|I|=S,k\in I \\ \delta(\Phi_I)>\delta_0}} \left[P(\psi_k) - P(\Psi_I)\right]\phi_k\right). \tag{104}$$

We further expand the sum over the well-conditioned sets using Sublemma B.9,

$$\begin{aligned} \sum_{\substack{|I|=S,k\in I \\ \delta(\Phi_I)\leq\delta_0}} \left[P(\psi_k) - P(\Psi_I)\right]\phi_k &= \sum_{\substack{|I|=S,k\in I \\ \delta(\Phi_I)\leq\delta_0}} (P(\Phi_I)b_k + \eta_{I,k}) \\ &= \sum_{\substack{|I|=S,k\in I \\ \delta(\Phi_I)\leq\delta_0}} (\Phi_I\Phi_I^\star b_k + \left[P(\Phi_I) - \Phi_I\Phi_I^\star\right] b_k + \eta_{I,k}) \\ &= \sum_{|I|=S,k\in I} \Phi_I\Phi_I^\star b_k - \sum_{\substack{|I|=S,k\in I \\ \delta(\Phi_I)>\delta_0}} \Phi_I\Phi_I^\star b_k + \sum_{\substack{|I|=S,k\in I \\ \delta(\Phi_I)\leq\delta_0}} (\left[P(\Phi_I) - \Phi_I\Phi_I^\star\right] b_k + \eta_{I,k}) \\ &= \binom{K-2}{S-2}\Phi\Phi^\star b_k - \sum_{\substack{|I|=S,k\in I \\ \delta(\Phi_I)>\delta_0}} \Phi_I\Phi_I^\star b_k + \sum_{\substack{|I|=S,k\in I \\ \delta(\Phi_I)\leq\delta_0}} (\left[P(\Phi_I) - \Phi_I\Phi_I^\star\right] b_k + \eta_{I,k}), \end{aligned} \tag{105}$$

where for the last equality we have used that $\langle b_k, \phi_k\rangle = 0$. Substituting the last expression into (104) we get,

$$\begin{aligned} \mathbb{E}(v) = \frac{C_r\gamma_{1,S}}{K}\Bigg[&\frac{S-1}{K-1}\Phi\Phi^\star b_k + \binom{K-1}{S-1}^{-1} \sum_{\substack{|I|=S,k\in I \\ \delta(\Phi_I)\leq\delta_0}} (\left[P(\Phi_I) - \Phi_I\Phi_I^\star\right] b_k + \eta_{I,k}) \\ &+ \binom{K-1}{S-1}^{-1} \sum_{\substack{|I|=S,k\in I \\ \delta(\Phi_I)>\delta_0}} \left(\left[P(\psi_k) - P(\Psi_I)\right]\phi_k - \Phi_I\Phi_I^\star b_k\right)\Bigg]. \end{aligned} \tag{106}$$

Substituting the bound $\|P(\Phi_I) - \Phi_I\Phi_I^\star\|_{2,2} \leq \delta(\Phi_I) \leq \delta_0$ as well as the bound for $\|\eta_{I,k}\|_2$ from Sublemma B.9 for the well-conditioned subsets and the bound

$$\left\| \left[P(\psi_k) - P(\Psi_I)\right]\phi_k \right\|_2 = \|P(\Psi_I)Q(\psi_k)\phi_k\|_2 \leq \|Q(\psi_k)\phi_k\|_2 = \sqrt{1 - |\langle \psi_k, \phi_k\rangle|^2} \leq \varepsilon_k \tag{107}$$

for the ill-conditioned subsets finally leads to

$$\begin{aligned}\|\mathbb{E}(v)\|_2 \leq \frac{C_r\gamma_{1,S}}{K} \Bigg[&\frac{S-1}{K-1}B\|b_k\|_2 + \delta_0\|b_k\|_2 + \frac{2\varepsilon\sqrt{S}}{\sqrt{(1-\delta_0)(1-\frac{\varepsilon^2}{2})} - 2\varepsilon\sqrt{S}} \cdot \|b_k\|_2 + \varepsilon\|b_k\|_2 \\ &+ \mathbb{P}(\delta(\Phi_I) > \delta_0 : |I| = S, k \in I) \cdot (\varepsilon_k + B\|b_k\|_2) \Bigg], \\ \leq \frac{C_r\gamma_{1,S}}{K} \Bigg[&\frac{SB}{K} + \delta_0 + \varepsilon + \frac{2\varepsilon\sqrt{S}}{\sqrt{(1-\delta_0)(1-\frac{\varepsilon^2}{2})} - 2\varepsilon\sqrt{S}} + (B+1)\mathbb{P}(\delta(\Phi_I) > \delta_0 : |I| = S, k \in I) \Bigg] \|b_k\|_2.\end{aligned} \tag{108}$$

If $\delta_S \leq \frac{1}{2}$, we choose $\delta_0 = \delta_S$, which for $S \leq \frac{K}{98B}$ and $\varepsilon \leq \frac{1}{32\sqrt{S}}$ leads to

$$\|\mathbb{E}(v)\|_2 \leq 0.611\varepsilon \cdot \frac{C_r\gamma_{1,S}}{K}. \tag{109}$$

In the non-trivial case, where Φ does not have a uniform isometry constant $\delta_S \leq \frac{1}{2}$, we can estimate (108) using J. Tropp's results on the conditioning of random subdictionaries. Reformulating Theorem 12 in [46] for our purposes we get that

$$\mathbb{P}(\delta(\Phi_I) > \delta_0 : |I| = S) \leq e^{-s} \qquad \text{for} \qquad s = \frac{\left(e^{-1/4}\delta_0 - \frac{2SB}{K}\right)^2}{144\mu^2 S}, \tag{110}$$

whenever $e^{-1/4}\delta_0 \geq \frac{2SB}{K}$, $s \geq \log(S/2+1)$ and $S \geq 4$. Together with the union bound,

$$\begin{aligned}\mathbb{P}(\delta(\Phi_I) > \delta_0 : |I| = S, k \in I) &= \binom{K-1}{S-1}^{-1} \sharp\{I : \delta(\Phi_I) > \delta_0, |I| = S, k \in I\} \\ &\leq \binom{K-1}{S-1}^{-1} \sharp\{I : \delta(\Phi_I) > \delta_0, |I| = S\} = \frac{K}{S} \cdot \mathbb{P}(\delta(\Phi_I) > \delta_0 : |I| = S),\end{aligned} \tag{111}$$

this leads to

$$\mathbb{P}(\delta(\Phi_I) > \delta_0 : |I| = S, k \in I) \leq \max\left\{S, \frac{K}{S}\right\} \exp\left(-\frac{\left(e^{-1/4}\delta_0 - \frac{2SB}{K}\right)^2}{144\mu^2 S}\right), \tag{112}$$

whenever $e^{-1/4}\delta_0 \geq \frac{2SB}{K}$ - in case one of the other original conditions is violated the statement is trivially true. Using the assumption $S \leq \frac{K}{98B}$, which does not represent a hard additional constraint, considering that in order to have $\varepsilon_{\mu,\rho} < 1$ we need $S \leq \frac{1}{98\mu^2}$ and that $\mu^2 \geq \frac{B-1}{K-1} \approx \frac{B}{K}$, we get for $\delta_0 = \frac{1}{4}$,

$$\mathbb{P}\left(\delta(\Phi_I) > \frac{1}{4} : |I| = S, k \in I\right) \leq K \exp\left(-\frac{1}{4741\mu^2 S}\right) := \varepsilon_\delta, \tag{113}$$

Substituting this bound for the choice $\delta_0 = \frac{1}{4}$ into (108) and using that $\varepsilon \leq \frac{1}{32\sqrt{S}}$ and $\varepsilon_\delta \leq \frac{1}{48(B+1)}$ we get

$$\|\mathbb{E}(v)\|_2 \leq 0.381\varepsilon \cdot \frac{C_r\gamma_{1,S}}{K}. \tag{114}$$

The second quantity we need to bound is the expected energy of $v = T(I,k)y \cdot \sigma(k) \cdot \chi(I,k)$,

$$\begin{aligned}
\mathbb{E}(\|v\|_2^2) &= \mathbb{E}_{c,p,\sigma,r}\left(\frac{\chi(I,k)}{1+\|r\|_2^2}\cdot\Big\|T(I,k)\Big(\sum_j \phi_j c\big(p(j)\big)\sigma(j)+r\Big)\Big\|_2^2\right)\\
&= \mathbb{E}_{c,p,r}\left(\frac{\chi(I,k)}{1+\|r\|_2^2}\left(\sum_j c\big(p(j)\big)^2\|T(I,k)\phi_j\|_2^2+\|T(I,k)r\|_2^2\right)\right)\\
&= \mathbb{E}_{p,r}\left(\frac{\chi(I,k)}{1+\|r\|_2^2}\left(\frac{\gamma_{2,S}}{S}\sum_{j\in I}\|T(I,k)\phi_j\|_2^2+\frac{1-\gamma_{2,S}}{K-S}\sum_{j\notin I}\|T(I,k)\phi_j\|_2^2+\|T(I,k)r\|_2^2\right)\right),\\
&\le \mathbb{E}_p\left(\chi(I,k)\left(\frac{\gamma_{2,S}}{S}\sum_{j\in I}\|T(I,k)\phi_j\|_2^2+\frac{1-\gamma_{2,S}}{K-S}\sum_{j\notin I}\|T(I,k)\phi_j\|_2^2+\mathbb{E}_r\left(\|T(I,k)r\|_2^2\right)\right)\right). \qquad (115)
\end{aligned}$$

We first estimate the two sums above given that $k \in I$. Note that we always have $\|P(\phi_k)-P(\psi_k)\|_{2,2} \le \varepsilon_k$ and $\|P(\phi_k)-P(\psi_k)\|_F \le \sqrt{2}\varepsilon_k$. Thus we get for the sum over I,

$$\begin{aligned}
\sum_{j\in I}\|T(I,k)\phi_j\|_2^2 &\le \sum_{j\in I}\left(\|[P(\Phi_I)-P(\Psi_I)]\phi_j\|_2+\|[P(\phi_k)-P(\psi_k)]\phi_j\|_2\right)^2\\
&= \sum_{j\in I}\left(\|Q(\Psi_I)]\phi_j\|_2+\|[P(\phi_k)-P(\psi_k)]\phi_j\|_2\right)^2\\
&\le \sum_{j\in I}\left(\|Q(\psi_j)]\phi_j\|_2+\|P(\phi_k)-P(\psi_k)\|_{2,2}\right)^2 \le \sum_{j\in I}\left(\varepsilon_j+\varepsilon_k\right)^2 \le 4S\varepsilon^2, \qquad (116)
\end{aligned}$$

and for the sum over the complement I^c,

$$\sum_{j\notin I}\|T(I,k)\phi_j\|_2^2 = \|T(I,k)\Phi_{I^c}\|_F^2 \le \|T(I,k)\|_F^2\|\Phi_{I^c}\|_{2,2}^2 \le B\|T(I,k)\|_F^2. \qquad (117)$$

To estimate the noise term in (115) we use the singular value decomposition of $T(I,k) = UDV^\star$,

$$\mathbb{E}\left(\|T(I,k)r\|_2^2\right) = \mathbb{E}\left(\|DV^\star r\|_2^2\right) = \mathbb{E}\left(\sum_i d_i^2|\langle v_i, r\rangle|^2\right) \le \sum_i d_i^2\rho^2 = \rho^2\|T(I,k)\|_F^2, \qquad (118)$$

where for the inequality we have used that for a subgaussian vector r with parameter ρ, the marginal $\langle v_i, r\rangle$ is subgaussian with parameter ρ. Substituting these estimates together with the bound $\|T(I,k)\|_F \le \|P(\Phi_I)-P(\Psi_I)\|_F+\sqrt{2}\varepsilon_k$ into (115) we get,

$$\mathbb{E}(\|v\|_2^2) \le \mathbb{E}_p\left(\chi(I,k)\left[4\gamma_{2,S}\varepsilon^2+\left(\frac{B(1-\gamma_{2,S})}{K-S}+\rho^2\right)\left(\|P(\Phi_I)-P(\Psi_I)\|_F+\sqrt{2}\varepsilon_k\right)^2\right]\right). \qquad (119)$$

As for the estimation of $\mathbb{E}(v)$ we now split the expectation over p into the well and the ill-conditioned subsets $I = p^{-1}(\mathbb{S})$. By Lemma A.2 in [39], whenever $\delta(\Phi_I) \le \delta_0$, we have

$$\|P(\Phi_I)-P(\Psi_I)\|_F^2 \le \frac{2\|Q(\Phi_I)B_I\|_F^2}{\sqrt{1-\delta_0}\left(\sqrt{1-\delta_0}-2\|B_I\|_F\right)} \qquad (120)$$

which for $\varepsilon \le \frac{1}{32\sqrt{S}}$ and $\delta_0 = 1/4$ (resp. $\delta_S \le 1/2$) simplifies to $\|P(\Phi_I)-P(\Psi_I)\|_F^2 \le 5S\varepsilon^2$. Together with

the general estimate $\|P(\Phi_I) - P(\Psi_I)\|_F \leq \sqrt{2S}$, this leads to

$$\begin{aligned}\mathbb{E}(\|v\|_2^2) &\leq \frac{S}{K}\left[4\gamma_{2,S}\varepsilon^2 + \left(\frac{B(1-\gamma_{2,S})}{K-S} + \rho^2\right)\left(\sqrt{5S}\varepsilon + \sqrt{2}\varepsilon_k\right)^2\right.\\ &\qquad \left. + \mathbb{P}\left(\delta(\Phi_I) > \frac{1}{4} : |I| = S, k \in I\right)\left(\frac{B(1-\gamma_{2,S})}{K-S} + \rho^2\right)\left(2S + 2\varepsilon_k\sqrt{S}\right)\right]\\ &\leq \frac{S}{K}\left[4\gamma_{2,S}\varepsilon^2 + 15\varepsilon^2\left(\frac{SB}{K-S}(1-\gamma_{2,S}) + S\rho^2\right)\right.\\ &\qquad \left. + \mathbb{P}\left(\delta(\Phi_I) > \frac{1}{4} : |I| = S, k \in I\right)\left(1 - \gamma_{2,S} + d\rho^2\right)\frac{2B(S+1)}{K-S}\right].\end{aligned}$$

Substituting the probability bound from (113) and assuming again that $S \leq \frac{K}{98B}$ as well as that $S \leq \frac{1}{98\rho^2}$ leads to the final estimate

$$\mathbb{E}(\|v\|_2^2) \leq \frac{S}{K}\left[5\varepsilon^2 + \frac{\varepsilon_\delta}{32}\left(1 - \gamma_{2,S} + d\rho^2\right)\right]. \tag{121}$$

Last we bound the norm of v in general as

$$\|v\|_2 = \|[P(\Phi_I) - P(\Psi_I) - P(\phi_k) + P(\psi_k)]y\|_2 \leq 2\|y\|_2 \leq 2\sqrt{B+1}. \tag{122}$$

In case $\gamma_{2,S} = 1, \rho = 0$ and therefore $y = \Phi_I x_I$ this reduces to

$$\begin{aligned}\|v\|_2 &\leq \|[\Phi_I - P(\Psi_I)\Phi_I\|_F\|x_I\|_2 + \|P(\phi_k) - P(\psi_k)\|_{2,2}\|\Phi_I x_I\|_2\\ &\leq \left(\sum_{i\in I}\|\phi_i - P(\Psi_I)\phi_i\|_2^2\right)^{\frac{1}{2}} + \varepsilon\sqrt{B} \leq \varepsilon\left(\sqrt{S} + \sqrt{B}\right),\end{aligned} \tag{123}$$

and in case of uniform isometry constant $\delta_S(\Phi) \leq 1/4$ and $\varepsilon \leq \frac{1}{32\sqrt{S}}$ to

$$\|v\|_2 \leq \|[P(\Phi_I) - P(\Psi_I)\|_F\|y\|_2 + \|P(\phi_k) - P(\psi_k)\|_{2,2}\|y\|_2 \leq \varepsilon\sqrt{B+1}\left(\sqrt{3S}+1\right). \tag{124}$$

Putting all the pieces together we get that under the assumptions in (a),

$$\begin{aligned}&\mathbb{P}\left(\frac{1}{N}\left\|\sum_n [R^o(\Psi, y_n, k) - R^o(\Phi, y_n, k)]\right\|_2 \geq \frac{C_r\gamma_{1,S}}{K}(0.381\varepsilon + t)\right)\\ &\qquad \leq \exp\left(-\frac{tC_r\gamma_{1,S}N}{8K}\min\left\{\frac{tC_r\gamma_{1,S}}{S\left[5\varepsilon^2 + \varepsilon_\delta\left(1 - \gamma_{2,S} + d\rho^2\right)/32\right]}, \frac{1}{3\sqrt{B+1}}\right\} + \frac{1}{4}\right)\\ &\qquad \leq \exp\left(-\frac{tC_r^2\gamma_{1,S}^2 N}{40K\max\{S, B+1\}}\min\left\{\frac{t}{\varepsilon^2 + \varepsilon_\delta\left(1 - \gamma_{2,S} + d\rho^2\right)/160}, \frac{3}{5}\right\} + \frac{1}{4}\right),\end{aligned}$$

under the assumptions in (b),

$$\begin{aligned}&\mathbb{P}\left(\frac{1}{N}\left\|\sum_n [R^o(\Psi, y_n, k) - R^o(\Phi, y_n, k)]\right\|_2 \geq \frac{C_r\gamma_{1,S}}{K}(0.381\varepsilon + t)\right)\\ &\qquad \leq \exp\left(-\frac{tC_r\gamma_{1,S}N}{8K}\min\left\{\frac{tC_r\gamma_{1,S}}{4\varepsilon^2 S}, \frac{1}{3\varepsilon\sqrt{S(B+1)}}\right\} + \frac{1}{4}\right)\\ &\qquad \leq \exp\left(-\frac{tC_r^2\gamma_{1,S}^2 N}{32\varepsilon K\max\{S, B+1\}}\min\left\{\frac{t}{\varepsilon}, 1\right\} + \frac{1}{4}\right),\end{aligned}$$

and under the assumptions in (c),

$$\begin{aligned}&\mathbb{P}\left(\frac{1}{N}\left\|\sum_n [R^o(\Psi, y_n, k) - R^o(\Phi, y_n, k)]\right\|_2 \geq \frac{\gamma_{1,S}}{K}(0.611\varepsilon + t)\right)\\ &\qquad \leq \exp\left(-\frac{t\gamma_{1,S}^2 N}{40\varepsilon K\max\{S, B+1\}}\min\left\{\frac{t}{\varepsilon}, 1\right\} + \frac{1}{4}\right).\end{aligned}$$

□

Remark B.4. For the lemma we have used both the sign and the permutation invariance, the sign invariance in (106) and the permutation invariance in (107). As for Lemma (B.5) but with a lot more effort, we can use the permutation invariance instead of using the sign invariance in (106). We will not go into details but via expanding the sum $T(I,k)\sum_{j\in I, j\neq k} x(j)\phi_j$, approximating $P(\Psi_I)\approx \Psi_I\Psi_I^\star$ and keeping track of how often an atom ϕ_j is in the support I one can show that as long as $S^2\lesssim K$ we still have $\|E(v)\|_2<\varepsilon\cdot C_r\gamma_{1,S}/K$ which is the necessary ingredient for the convergence proof. An alternative criterion, that trades off permutation invariance for sign invariance, is again the one discussed in Remark B.2. However it is not enough to preserve Eq. (107), where we need that $\|\mathbb{E}_{I:k\in I}\Phi^I\Phi_I^\star b_k\|_2\leq\varepsilon$. For this inequality we do not only need to avoid that two atoms ϕ_j and ϕ_k are always used in the same ratio, but that they are always used together no matter the ratio, because any two atoms $\tilde\phi_j$ and $\tilde\phi_k$ which span the same subspace have the same approximation properties. Indeed if $x(j)$ and $x(k)$ are both randomly $\pm1/\sqrt{S}$ then $\tilde\phi_j=\phi_j+\phi_k$ and $\tilde\phi_k=\phi_j-\phi_k$ actually provide sparser approximations.

Sublemma B.9. *Let Φ_I be a subdictionary of Φ with $\delta(\Phi_I)\leq\delta_0$ and Ψ_I the corresponding subdictionary of an ε-perturbation of Ψ, that is $d(\Phi,\Psi)=\varepsilon$. If $k\in I$ then*

$$\big[P(\psi_k)-P(\Psi_I)\big]\phi_k=P(\Phi_I)b_k+\eta_{I,k}\quad\text{ with }\quad\|\eta_{I,k}\|_2\leq\left(\frac{2\varepsilon\sqrt{S}}{\sqrt{(1-\delta_0)(1-\frac{\varepsilon^2}{2})}-2\varepsilon\sqrt{S}}+\varepsilon\right)\cdot\|b_k\|_2. \tag{125}$$

Proof: If $\delta(\Phi_I)\leq\delta_0$ we can use the expression for $P(\Psi_I)$ developed in Lemma A.2 of [39],

$$P(\Psi_I)=\big(\Phi_I+Q(\Phi_I)B_IM_I\big)(\Phi_I^\star\Phi_I)^{-1}\left(\mathbb{I}_S+\sum_{i=1}^\infty(-R_I)^i\right)\big(\Phi_I+Q(\Phi_I)B_IM_I\big)^\star,$$
$$\text{with}\qquad M_I=\mathbb{I}_S+\sum_{i=1}^\infty(-\Phi_I^\dagger B_I)^i\qquad\text{and}\qquad R_I=M_I^\star B_I^\star Q(\Phi_I)B_IM_I(\Phi_I^\star\Phi_I)^{-1} \tag{126}$$

to get $P(\psi_k)\phi_k=\alpha_k^2(\phi_k+b_k)$ and

$$\begin{aligned}
P(\Psi_I)\phi_k&=\big(\Phi_I+Q(\Phi_I)B_IM_I\big)(\Phi_I^\star\Phi_I)^{-1}\left(\mathbb{I}_S+\sum_{i=1}^\infty(-R_I)^i\right)\Phi_I^\star\phi_k\\
&=\phi_k+Q(\Phi_I)B_IM_I(\Phi_I^\star\Phi_I)^{-1}\Phi_I^\star\phi_k+\big(\Phi_I+Q(\Phi_I)B_IM_I\big)(\Phi_I^\star\Phi_I)^{-1}\sum_{i=1}^\infty(-R_I)^i\Phi_I^\star\phi_k\\
&=\phi_k+Q(\Phi_I)B_I\left(\mathbb{I}_S+\sum_{i=1}^\infty(-\Phi_I^\dagger B_I)^i\right)e_{k|I}+\big(\Phi_I+Q(\Phi_I)B_IM_I\big)(\Phi_I^\star\Phi_I)^{-1}\sum_{i=1}^\infty(-R_I)^i\Phi_I^\star\phi_k\\
&=\phi_k+b_k-P(\Phi_I)b_k+Q(\Phi_I)B_I\sum_{i=1}^\infty(-\Phi_I^\dagger B_I)^ie_{k|I}+\big(\Phi_I+Q(\Phi_I)B_IM_I\big)(\Phi_I^\star\Phi_I)^{-1}\sum_{i=1}^\infty(-R_I)^i\Phi_I^\star\phi_k.
\end{aligned}$$

Subtracting the projections we see that all that remains to do is to estimate the size of

$$\eta_{I,k}:=Q(\Phi_I)B_IM_I(\Phi_I^\dagger B_I)e_{k|I}-\big((\Phi_I^\dagger)^\star+Q(\Phi_I)B_IM_I(\Phi_I^\star\Phi_I)^{-1}\big)\sum_{i=1}^\infty(-R_I)^i\Phi_I^\star\phi_k-\frac{\omega_k^2}{\alpha_k}\psi_k. \tag{127}$$

Using standard bounds for matrix vector products and the identity $\|(\Phi_I^\star\Phi_I)^{-1}\|_{2,2}=\|\Phi_I^\dagger\|_{2,2}^2$ we get

$$\begin{aligned}
\|\eta_{I,k}\|_2&\leq\|B_IM_I\|_{2,2}\|\Phi_I^\dagger b_k\|_2+\Big(\|\Phi_I^\dagger\|_{2,2}+\|B_IM_I\|_{2,2}\|\Phi_I^\dagger\|_{2,2}^2\Big)\sum_{i=0}^\infty\|R_I\|_{2,2}^i\|R_I\Phi_I^\star\phi_k\|_2+\frac{\omega_k^2}{\alpha_k}\\
&\leq\|B_IM_I\|_{2,2}\|\Phi_I^\dagger\|_{2,2}\|b_k\|_2+\Big(\|\Phi_I^\dagger\|_{2,2}+\|B_IM_I\|_{2,2}\|\Phi_I^\dagger\|_{2,2}^2\Big)\sum_{i=0}^\infty\Big(\|\Phi_I^\dagger\|_{2,2}^2\|B_IM_I\|_{2,2}^2\Big)^i\|R_I\Phi_I^\star\phi_k\|_2+\frac{\omega_k^2}{\alpha_k}.
\end{aligned}$$

We next expand $R_I\Phi_I^\star\phi_k$ remembering the definition of R_I and M_I as

$$\begin{aligned}R_I\Phi_I^\star\phi_k &= M_I^\star B_I^\star Q(\Phi_I)B_I\left(\mathbb{I}_S+\sum_{i=1}^{\infty}(-\Phi_I^\dagger B_I)^i\right)(\Phi_I^\star\Phi_I)^{-1}\Phi_I^\star\phi_k\\ &= M_I^\star B_I^\star Q(\Phi_I)\left(\mathbb{I}_d+\sum_{i=1}^{\infty}(-B_I\Phi_I^\dagger)^i\right)B_Ie_{k|I} = M_I^\star B_I^\star Q(\Phi_I)\left(\mathbb{I}_d+\sum_{i=1}^{\infty}(-B_I\Phi_I^\dagger)^i\right)b_k\end{aligned}$$

to get

$$\|R_I\Phi_I^\star\phi_k\|_2 \le \|B_IM_I\|_{2,2}\left(1-\|B_I\|_{2,2}\|\Phi_I^\dagger\|_{2,2}\right)^{-1}\|b_k\|_2.$$

Substituting this estimate together with the bound $\|M_I\|_{2,2} \le \left(1-\|B_I\|_{2,2}\|\Phi_I^\dagger\|_{2,2}\right)^{-1}$ into the above bound for $\|\eta_{I,k}\|_2$, resolving the sums and fractions and noting that $\|b_k\|_2 = \frac{\omega_k}{\alpha_k}$ leads to,

$$\|\eta_{I,k}\|_2 \le \left(\frac{2\|B_I\|_{2,2}}{\|\Phi_I^\dagger\|_{2,2}^{-1}-2\|B_I\|_{2,2}}+\omega_k\right)\cdot\|b_k\|_2.$$

To get to the final statement we use the bounds $\|B_I\|_{2,2}^2 \le \|B_I\|_F^2 \le S\varepsilon^2/(1-\varepsilon^2/2)$ and $\|\Phi_I^\dagger\|_{2,2}^{-1} \ge \sqrt{1-\delta(\Phi_I)} \ge \sqrt{1-\delta_0}$. □

Lemma B.10. *If for two vectors ψ,ϕ, where $\|\phi\|_2=1$, and two scalars $0<t<s$ we have, $\|\psi-s\phi\|_2^2 \le t^2$ then*

$$\left\|\frac{\psi}{\|\psi\|_2}-\phi\right\|_2^2 \le 2-2\sqrt{1-\frac{t^2}{s^2}}. \tag{128}$$

Proof: Writing $\psi=\alpha\phi+\omega z$ for some unit norm vector z with $\langle z,\phi\rangle=0$ we can reformulate the initial constraint $\|\psi-s\phi\|_2^2\le t^2$ to $(\alpha-s)^2+\omega^2\le t^2$, while the quantity whose maximal size we have to estimate becomes

$$\left\|\frac{\psi}{\|\psi\|_2}-\phi\right\|_2^2 = 2-2\frac{\alpha}{\sqrt{\alpha^2+\omega^2}}. \tag{129}$$

Solving the resulting maximisation problem we get that the maximum is attained at $\alpha=\frac{s^2-t^2}{s}$ and $\omega=\frac{t}{s}\sqrt{s^2-t^2}$ and that therefore

$$\left\|\frac{\psi}{\|\psi\|_2}-\phi\right\|_2^2 \le 2-2\sqrt{1-\frac{t^2}{s^2}}. \tag{130}$$

□

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, A. Bhaskara, R. Ge, and T. Ma. More algorithms for provable dictionary learning. *arXiv:1401.0579*, 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] S. Arora, R. Ge, and A. Moitra. New algorithms for learning incoherent and overcomplete dictionaries. In *COLT 2014 (arXiv:1308.6273)*, 2014.

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

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

[9] T. Blumensath and M.E. Davies. Iterative thresholding for sparse approximations. *Journal of Fourier Analysis and Applications*, 14(5-6):629–654, 2008.

[10] T. Blumensath and M.E. Davies. Iterative Hard Thresholding for compressed sensing. *Applied Computational Harmonic Analysis*, 27(3):265–274, 2009.

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

[12] O. Christensen. *An Introduction to Frames and Riesz Bases*. Birkhäuser, 2003.

[13] 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.

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

[15] 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.

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

[17] Q. Geng, H. Wang, and J. Wright. On the local correctness of ℓ^1-minimization for dictionary learning. *arXiv:1101.5672*, 2011.

[18] P. Georgiev, F.J. Theis, and A. Cichocki. Sparse component analysis and blind source separation of underdetermined mixtures. *IEEE Transactions on Neural Networks*, 16(4):992–996, 2005.

[19] 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.

[20] R. Gribonval, R. Jenatton, F. Bach, M. Kleinsteuber, and M. Seibert. Sample complexity of dictionary learning and other matrix factorizations. *IEEE Transactions on Information Theory*, 61(6):3469–3486, 2015.

[21] 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.

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

[23] E. Höck. Hard thresholding pursuit for sparse approximation. BSc thesis, University of Innsbruck, 2016.

[24] P. Jain, A. Tewari, and P. Kar. On iterative hard thresholding methods for high-dimensional m-estimation. In *NIPS14 (arXiv:14105137)*, 2014.

[25] A. Jung, Y. Eldar, and N. Görtz. Performance limits of dictionary learning for sparse coding. In *EUSIPCO14 (arXiv:1402.4078)*, pages 765 – 769, 2014.

[26] 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.

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

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

[29] M. Ledoux and M. Talagrand. *Probability in Banach spaces. Isoperimetry and processes.* Springer-Verlag, Berlin, Heidelberg, NewYork, 1991.

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

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

[32] 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.

[33] 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.

[34] A. Maurer and M. Pontil. K-dimensional coding schemes in Hilbert spaces. *IEEE Transactions on Information Theory*, 56(11):5839–5846, 2010.

[35] N.A. Mehta and A.G. Gray. On the sample complexity of predictive sparse coding. *arXiv:1202.4050*, 2012.

[36] V. Naumova and K. Schnass. Dictionary learning from incomplete data, Part I algorithms. *in preparation*, 2016.

[37] M.D. Plumbley. Dictionary learning for ℓ_1-exact sparse coding. In M.E. Davies, C.J. James, and S.A. Abdallah, editors, *International Conference on Independent Component Analysis and Signal Separation*, volume 4666, pages 406–413. Springer, 2007.

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

[39] K. Schnass. On the identifiability of overcomplete dictionaries via the minimisation principle underlying K-SVD. *Applied Computational Harmonic Analysis*, 37(3):464–491, 2014.

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

[41] K. Schnass. Sequential dictionary learning with model selection. *in preparation*, 2016.

[42] K. Schnass and P. Vandergheynst. Average performance analysis for thresholding. *IEEE Signal Processing Letters*, 14(11):828–831, 2007.

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

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

[45] J. Sun, Q. Qu, and J. Wright. Complete dictionary recovery over the sphere. In *ICML 2015 (arXiv:1504.06785)*, 2015.

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

[47] D. Vainsencher, S. Mannor, and A.M. Bruckstein. The sample complexity of dictionary learning. *Journal of Machine Learning Research*, 12(3259-3281), 2011.
[48] M. Yaghoobi, T. Blumensath, and M.E. Davies. Dictionary learning for sparse approximations with the majorization method. *IEEE Transactions on Signal Processing*, 57(6):2178–2191, June 2009.
[49] M. Zibulevsky and B.A. Pearlmutter. Blind source separation by sparse decomposition in a signal dictionary. *Neural Computations*, 13(4):863–882, 2001.