Compressed Sensing and Dictionary Learning with non-uniform support distribution

Dissertation in Mathematics

Submitted by

Simon Ruetz

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

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

Advisor: Univ.-Prof. Dr. Karin Schnass

Innsbruck, 17th August 2022

Abstract

With ever growing amounts of data, the task of finding and using underlying structure is becoming more and more important. One influential insight, which modern signal processing tasks like compressed sensing or dictionary learning rely on, is that many data types allow some kind of sparse representation. So given a suitable basis, a small number of elements is sufficient to approximate any given signal.
In the first part of this thesis we derive technical results allowing us to analyse compressed sensing and dictionary learning in more general settings than previously possible. Concretely, we derive tail bounds for the operator norm of random submatrices with non-uniformly distributed supports, essentially showing that for a well-conditioned matrix most submatrices behave like an isometry.
Compressed sensing consists of reconstructing a sparse signal from a small number of linear measurements. If these measurements are chosen from a bigger set of possible linear measurements via sampling from a (possibly non-uniform) probability distribution, this is called (variable density) subsampling. In the next part of the thesis we derive an optimal subsampling density, guaranteeing recovery in a very general setting. More precisely, we show how the optimal subsampling density depends on the structure of the sensing matrix and on the distribution of the sparse supports, which can easily be estimated from data. This leads to a simple formula for subsampling densities achieving state of the art performance in numerical experiments. Our approach also extends to structured acquisition, where instead of isolated measurements, blocks of measurements are taken.
In some applications the sparsifying basis is not known or one wants to find a better one. Learning such a basis from data is called dictionary learning. In the third part of this thesis we study the convergence behaviour of two of the most popular dictionary learning algorithms - the Method of Optimal Directions (MOD) and the Approximate K-SVD (aK-SVD). By again using our bounds for operator norms of non-uniformly distributed submatrices, we are able to use a very general non-uniform signal model and derive sufficient conditions for the convergence of these algorithms, improving greatly upon existing results.
In the last part we analyse the performance of two sparse approximation algorithms, Orthogonal Matching Pursuit (OMP) and Thresholding in the case in which only a perturbed version of the basis is known. Both theory and numerical simulations show that the computationally lighter Thresholding algorithm is a viable alternative to OMP in applications such as dictionary learning.

Acknowledgements

I want to use this opportunity to wholeheartedly thank my supervisor Dr. Karin Schnass for her continued support, her encouragement and all the fun coffee breaks throughout this PhD. I hope that her crusade against my — admittedly rather dense — notational style, which undoubtedly cost her a lot of nerves, bore fruit in this thesis.

I also want to thank my family, friends and my girlfriend for providing me with continued love and support. I could not have done it without you.

I also thank the Austrian Science Fund (FWF) for the employment under the Grant no. Y760.

Contents

1 Introduction **9**
1.1 Outline . . . 11
1.2 Notation . . . 12

2 Submatrices with non-uniformly random supports **15**
2.1 Introduction . . . 15
2.2 Main results . . . 17
2.3 Application to sparse approximation . . . 22
2.4 Sensing dictionaries and preconditioning . . . 27
2.5 Proof of operator norm concentration . . . 29
2.6 Sensing matrices . . . 35
2.7 Discussion . . . 35

3 Adapted variable density subsampling **37**
3.1 Compressed Sensing . . . 37
3.2 Contribution . . . 38
3.3 Main result . . . 38
3.4 Special cases . . . 40
3.5 Numerical experiments . . . 42
3.6 Sparsity in levels and blocks of measurements . . . 44
3.7 Proof of Theorem 3.2 . . . 48
3.8 Discussion . . . 52

4 Dictionary learning convergence **53**
4.1 Introduction . . . 53
4.2 Setting . . . 55
4.3 Algorithms . . . 55
4.4 Main results . . . 60
4.5 Proof . . . 62
4.6 Technical results . . . 69
4.7 Sparse approximation and conditioning of subdictionaries . . . 71

4.8 Proof of Claims 1-4 73
4.9 How to calculate expectations in the rejective sampling model 85
4.10 Discussion 96

5 OMP vs Thresholding under dictionary mismatch 97
5.1 Introduction 97
5.2 Setting 99
5.3 Main results 99
5.4 Comparison of OMP and Thresholding 101
5.5 Dictionary learning using OMP and Thresholding 102
5.6 Discussion 105

6 Discussion and Outlook 107

Chapter 1

Introduction

Sparse representations are crucial for modern signal processing tasks like denoising, compression, compressed sensing, inpainting and many more. When talking about sparsity, we usually mean that a signal is either directly sparse, i.e. has only a small number of non-zero entries, or has a sparse representation in a suitable basis or overcomplete system, meaning that it can be written as a linear combination of a small number of elements.
Sparsity for example lies at the heart of modern compression algorithms like the JPEG and JPEG2000 standards which transform images into a different basis before using a threshold to keep only the largest coefficients, thus saving space when storing data. Frequently used transforms (especially for image data) are the discrete cosine transform (DCT) or some wavelet transform, for which it is well known that the resulting representations are approximately sparse. These two transforms correspond to orthogonal bases, which are very easy to handle. Unfortunately, in many applications we have to resort to overcomplete systems, called dictionaries, to get decent sparse representations for a particular signal class.
Formally, we represent a signal $y \in \mathbb{R}^d$ as a linear combination of $S \ll d$ elements of $\Phi = (\phi_1, \cdots, \phi_K) \in \mathbb{R}^{d \times K}$ with $d < K$, via

$$y = \sum_{i \in I} \phi_i x_i = \Phi_I x_I \quad \text{s.t.} \quad |I| = S,$$

where Φ_I denotes the restriction to the columns indexed by I, called the support.
Nevertheless, such a representation in a dictionary also has some drawbacks. For example finding the vector x and its support I given the dictionary Φ and signal y is combinatorial in nature. So to avoid searching through all possible sets I, suboptimal routines are typically used. This is called sparse approximation and some of the most popular sparse approximation algorithms are Orthogonal Matching Pursuit (OMP), Thresholding or Basis Pursuit (BP). Deriving sufficient conditions for these algorithms to successfully recover x and I given the dictionary Φ and signal y will be a big part in the first part of the thesis. We will see that success of these algorithms depends to a large extend on bounding the extreme singular values of Ψ_I.
In the next part we will analyse a close cousin of sparse approximation, compressed sensing. There, we usually talk about a measurement matrix A which, contrary to using a fixed dictionary Φ, has some design choices. We try to recover an unknown signal from as few measurements as possible, where by measurements we mean a linear map $A \in \mathbb{R}^{m \times d}$, taking a signal $x \in \mathbb{R}^d$ and mapping it to a vector $y \in \mathbb{R}^m$ via $y = Ax$. As we want to minimise the number of measurements, we usually have $m \ll d$ and thus the above system of linear equations

is underdetermined, meaning there is either no solution or infinitely many. So by assuming that x is sparse (or sparse in a orthogonal basis) we hope to get uniqueness by searching for the sparsest vector, such that the linear equations are satisfied. So similar in spirit as sparse approximation we want to solve

$$\hat{x} = \arg\min \|x\|_0 \quad \text{s.t.} \quad y = Ax.$$

As this is non-convex and in general NP-hard, compressed sensing traditionally focuses on the convex relaxation of this minimisation problem — replacing the ℓ_0-norm by the ℓ_1-norm and instead trying to solve

$$\hat{x} = \arg\min \|x\|_1 \quad \text{s.t.} \quad y = Ax.$$

Finding conditions on the matrix A and lower bounds on the number of measurements m such that the above minimisation problem recovers x with a given sparsity level S are some of the main research goals in the vast field of compressed sensing [17, 32]. Some of the best theoretical lower bounds on m in terms of the sparsity level S and signal dimension d are achieved by using either a Gaussian or Bernoulli random matrix. While such random designs also yield very good practical performance, in applications like Computed Tomography (CT) or Magnetic Resonance Imaging (MRI) such random measurements are not possible due to existing hardware constraints.

These applications inspired the setup that we will look at in this thesis, where the sensing matrix $A \in \mathbb{C}^{m\times K}$ is constructed by choosing m rows from some bigger matrix $A_0 \in \mathbb{C}^{K\times K}$ which represents the set of possible linear measurements. The question then becomes how to best spend this budget of m rows, i.e. how to pick the "best" linear measurements from this set of possible linear measurements. Usually these strategies are characterised via a (possibly non-uniform) probability distribution that tells us which row of A_0 should be picked with which probability. This setting is called (variable density) subsampling and we will analyse it in the next part of the thesis.

All of the above applications assume knowledge of a basis Φ in which our given signals have a sparse representation. Yet not for all signal classes such a basis is known, or in some cases we would like to have a better one. For the signal class of images it is for example well known that they can be sparsely represented in the 2D DCT basis, yet when dealing only with a certain subclass (think of images of trees or knees), we would like to find a better dictionary that is even more adapted to the data we want to represent. This process of learning a dictionary from data is called dictionary learning. Formally we try to decompose a data matrix $Y \in \mathbb{R}^{d\times N}$, where each column corresponds to one signal, into a dictionary $\Phi \in \mathbb{R}^{d\times K}$ and a sparse coefficient matrix $X \in \mathbb{R}^{K\times N}$ such that

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

There exist many algorithms to try and solve this problem. The most popular among them belong to the class of alternating minimisation algorithms, which alternate between updating the dictionary Φ, while keeping the sparse coefficients X fixed, and vice versa. The next part of this thesis will analyse the convergence behaviour of two of the most popular such alternating minimisation dictionary learning algorithms - the Method of Optimal Directions (MOD) and a variation of the Approximate K-SVD (aK-SVD). When analysing these algorithms, one big hurdle is to control the sparse approximation step between the dictionary updates. Since alternating minimisation algorithms rely heavily on the choice of sparse approximation algorithm, we finish this thesis by arguing that Thresholding is a viable alternative to OMP in settings where only a perturbed version of the generating dictionary is available. This is the reason why we analyse MOD and K-SVD with Thresholding as sparse approximation algorithm instead of the computationally more complex OMP algorithm.

1.1. Outline

This thesis will be structured as follows. Section 1.2 will define the setting and commonly used notation.

In Chapter 2 we will see that in order to get theoretical results for sparse approximation algorithms that reflect practical performance, we have to control the operator norms of the submatrices Φ_I. Deterministic bounds on $\|\Phi_I^*\Phi_I - \mathbb{I}\|_{2,2}$ for *arbitrary* supports I are, unfortunately, only of limited use, since they are very restrictive with regards to the matrix Φ. This led to the emergence of results where the support I is chosen uniformly at random amongst all possible subsets with cardinality S, thus deriving bounds on $\|\Phi_I^*\Phi_I - \mathbb{I}\|_{2,2}$ for *most* supports I. These early results yielded powerful concentration inequalities for the operator norms of these random submatrices yet they rely on the assumption that the support I is chosen uniformly at random. Unfortunately, to model more general signal classes we show that it is necessary to analyse the operator norms of non-uniformly selected random submatrices. Thus in Chapter 2 we will derive concentration inequalities for the operator norms of non-uniformly selected submatrices. These results, for instance, allow us to conduct an analysis of the average case performance of Thresholding, Orthogonal Matching Pursuit and Basis Pursuit under very general model assumptions.

In Chapter 3 we will turn to compressed sensing in a subsampling setup, where the sensing matrix A is constructed by sampling rows form a bigger matrix A_0 comprising all possible linear measurements. Whereas early results in compressed sensing theory focused mainly on uniform sampling methods, more recent results showed that, apart from the structure of the matrix A_0, the optimal subsampling strategy should also take the sparsity pattern of the signal into account. The obvious caveat of these results is that in most practical application this oracle-like knowledge remains elusive.
We close this gap by characterising the structured sparsity via a probability distribution p on the supports of the sparse signals, allowing us to again derive optimal subsampling strategies. Given access to a dataset of similar signals, the probability distribution p can be easily estimated and we show that this technique achieves state of the art performance in numerical experiments. In practice, instead of isolated measurements (single rows of A_0), often blocks of measurements are taken. We extend our results to this setting and show how to again derive an optimal sampling strategy. Once again, our knowledge of how to bound submatrices with non-uniformly distributed sparse supports is the key to the theoretical results in this chapter.

In Chapter 4 we will derive sufficient conditions for the dictionary learning algorithms MOD and aK-SVD to recover an underlying ground truth under much more relaxed conditions than previously thought necessary. We will show that if the distance $\max_i \|\psi_i - \phi_i\|_2$ between the generating dictionary Φ, i.e. the ground truth, and the initialisation Ψ is smaller than $1/\log(K)$, then both dictionary learning algorithms with Thresholding as the sparse approximation algorithm will recover the generating dictionary. This in itself is already a huge improvement upon existing results, but the true strength of our result is to expand this radius of convergence to distances close to $\sqrt{2}$, if Φ and Ψ exhibit a certain structure, where each element in Ψ only points to one element of Φ. This confirms intuition that even for very large distances between the initialisation and the generating dictionary, as long as there is an obvious choice for each element in the initialisation to converge to, it will find the solution. We will further use a very

general signal model with non-uniformly distributed sparse supports by again using the results about the operator norms of random submatrices derived in Chapter 2.

In Chapter 5 we study the performance of OMP in comparison to Thresholding in the case in which only a perturbed version of the generating dictionary is known. Both theory and numerical simulations show that even for reasonably small perturbations, the advantage of the computationally more complex OMP algorithm disappears. Moreover, simulations also indicate that Thresholding is better at avoiding spurious local minima in the optimisation landscape of dictionary learning. This is the reason why we only analysed Thresholding in the sparse approximation step in the dictionary learning algorithms in the previous chapter, even thought the original MOD and K-SVD algorithms used OMP as sparse approximation algorithm.

Finally, Chapter 6 finishes this thesis with an outlook on possible future research directions and a discussion of open problems.

1.2. Notation

A quick note on the notation used throughout this text. For an integer K, we write $\mathbb{K} := \{1, \cdots, K\}$. The vectors $(e_i)_{1 \leq i \leq K}$ denote the vectors of the canonical basis of $\mathbb{R}^K$. For a matrix $A \in \mathbb{C}^{d \times K}$, we denote by $A_{:,k}$ (resp. $A_{k,:}$) the k-th column (resp. row) of A and by $A_{J,L}$ the submatrix with rows indexed by set $J \subseteq \{1, \cdots, d\}$ and columns indexed by set $L \subseteq \mathbb{K}$. If we talk about certain columns of a matrix A, we often drop the second index, i.e. instead of $A_{:,k}$ we will write A_k and instead of $A_{:,J}$, we will write A_J. By A^* we denote the conjugate transpose of the matrix A and by $A_k^* \in \mathbb{R}^{1 \times d}$, the conjugate transpose of the k-th column of A. We denote by $A_J^\dagger$ the Moore-Penrose pseudoinverse of the matrix A_J and by $P(A_J) := A_J A_J^\dagger$ the projection onto the column span of A_J.
For $1 \leq p, q, r \leq \infty$ we set

$$\|A\|_{p,q} := \max_{\|x\|_q = 1} \|Ax\|_p.$$

So for $B \in \mathbb{C}^{K \times m}$ we get $\|AB\|_{p,q} \leq \|A\|_{q,r} \|B\|_{r,p}$ and $\|Ax\|_q \leq \|A\|_{q,p} \|x\|_p$. Frequently encountered quantities are

$$\|A\|_{\infty,2} = \max_{k \in \{1,\ldots,d\}} \|A^k\|_2 \quad \text{and} \quad \|A\|_{2,1} = \max_{k \in \{1,\ldots,K\}} \|A_k\|_2,$$

which denote the maximum ℓ_2-norm of a row and the maximum ℓ_2-norm of a column of A respectively. Note that $\|A\|_{\infty,2} = \|A^*\|_{2,1}$. Further note that $\|A\|_{\infty,1}$ simply is the maximum absolute entry of the matrix A. For ease of notation we sometimes write $\|A\| = \|A\|_{2,2}$ for the operator norm which corresponds to the largest singular value of A. For a vector $v \in \mathbb{R}^d$, we denote by $\underline{v} := \|v\|_{\min} := \min_i |v_i|$ the smallest absolute entry of v and $\|v\|_{\max} := \|v\|_\infty$ the biggest absolute entry of v. We write $x \lesssim y$ if there exists a constant $c > 0$, such that $x \leq cy$. We write $\mathrm{vec} : \mathbb{C}^{d \times d} \mapsto \mathbb{C}^{d^2}$ for the vectorisaton operation that transforms a complex matrix into a complex vector by stacking the columns on top of each other and by vec^{-1} its inverse. Further, for any vector v we denote by D_v resp. $\mathrm{diag}(v)$ the diagonal matrix with v on the diagonal and abbreviate $D_{v \cdot w} := D_v \cdot D_w$. Finally we write $\odot$ for the Hadamard Product (or pointwise product) of two matrices of the same dimension.
In the following we will often consider dictionaries Φ and Ψ, i.e., a collection of K unit norm vectors $\phi_i, \psi_i \in \mathbb{R}^d$, called atoms, and define the coherence of a dictionary Φ as $\mu(\Phi) :=$

$\max_{i\neq j} |\langle \phi_i, \phi_j \rangle|$. For a dictionary Φ and an index set I of size S we define $\vartheta(\Phi_I) = \|\Phi_I^* \Phi_I - \mathbb{I}_S\|_{2,2}$. Note that if $\vartheta(\Phi_I) < 1$ and therefore Φ_I has full rank, we have for the projection $P(\Phi_I) = \Phi_I(\Phi_I^*\Phi_I)^{-1}\Phi_I^*$. If it is clear from context, we will write ϑ instead of $\vartheta(\Phi_I)$ and μ instead of $\mu(\Phi)$.
As was noted in the introduction we want the supports to follow a non-uniform distribution, allowing some columns, to be picked more frequently than others. For a set $\mathbb{K}$ we denote by $\mathcal{P}(\mathbb{K})$ the power set (set of all subsets) of $\mathbb{K}$. We are going to use the following two sampling models which define two discrete probability measures on $\mathcal{P}(\mathbb{K})$ that allow us to model non-uniform distributions for our supports.

Definition 1.1 (Poisson sampling) *Let δ_j denote a sequence of K independent Bernoulli* 0-1 *random variables with expectations p_j such that $\sum_{j=1}^K p_j = S$. We say the support I follows the Poisson sampling model, if $I := \{i \mid \delta_i = 1\}$. Each support $I \subseteq \mathbb{K}$ is chosen with probability*

$$\mathbb{P}_B(I) = \prod_{i\in I} p_i \prod_{j\notin I}(1-p_j). \tag{1.1}$$

Supports following a Poisson sampling model have (by definition of the Bernoulli r.v.) cardinality S *on average*. This comes with the big advantage that the probability of one atom appearing in the support is independent of the others, allowing us to use concentration inequalities for sums of independent random matrices later on. The drawback of this model is that the supports are not exactly S sparse. This can be achieved by keeping only those supports that have cardinality S and *throwing away* the rest. This amounts to simply conditioning the above Poisson sampling model on the event that exactly S of the Bernoulli r.v. are equal to 1, leading to our second support distribution model.

Definition 1.2 (Rejective sampling - Conditional Bernoulli) *Let δ_j denote a sequence of K independent Bernoulli* 0-1 *random variables with expectations p_j such that $\sum_{j=1}^K p_j = S$ and denote by $\mathbb{P}$ the probability measure of the corresponding Poisson sampling model. We say our support I follows the rejective sampling model, if each support $I \subseteq \mathbb{K}$ is chosen with probability*

$$\mathbb{P}_S(I) := \mathbb{P}_B(I \mid |I| = S) = \begin{cases} c\prod_{i\in I} p_i \prod_{j\notin I}(1-p_j) & \text{if} \quad |I| = S \\ 0 & \text{else} \end{cases}, \tag{1.2}$$

where c is a constant to ensure that $\mathbb{P}_S$ is a probability measure.

The distributions of the supports in the above two sampling models are uniquely defined by the expectations of the Bernoulli random variables p_ℓ. For more information on Poisson and rejective sampling, we refer the interested reader to [42]. We define the square diagonal matrix $D_{\sqrt{p}} := \operatorname{diag}((\sqrt{p_\ell})_\ell)$ and set the inclusion probabilities in the rejective sampling model $\pi_\ell := \mathbb{P}_S(\ell \in I)$. Note that $\pi_\ell \neq p_\ell$ except for the case where $\pi_\ell = S/K$ for all ℓ. This is a result of the rejective sampling model with its intrinsic dependence. The square diagonal matrix $D_{\sqrt{\pi}} := \operatorname{diag}((\sqrt{\pi_\ell})_\ell)$ will also be used often throughout this thesis. We further set $R_I := (\mathbb{I}_I)^* \in \mathbb{R}^{|I|\times K}$, allowing us to write $A_I = AR_I^*$. This also allows us to embed a matrix $A_I \in \mathbb{R}^{d\times S}$ into $\mathbb{R}^{d\times K}$ by zero-padding via $A_I R_I \in \mathbb{R}^{d\times K}$. We denote by $\mathbf{1}_I \in \mathbb{R}^K$ the vector, whose entries indexed by I are 1 and zero else.
Let D_I be the square diagonal *selector matrix* whose diagonal entries are the δ_ℓ, i.e., set $D_I := \operatorname{diag}((\delta_\ell)_\ell)$. Note that there is a one to one correspondence between selector matrices

and index sets, i.e., $(D_I)_{\ell,\ell} = 1 \Leftrightarrow \ell \in I$. Thus, any probability measure on the index sets I induces a probability measure on the set of selector matrices. Note also that we have

$$D_I = R_I^* R_I = \mathrm{diag}(\mathbf{1}_I).$$

Further, for $A \in \mathbb{R}^{K\times K}$ and $j,k \in \mathbb{K}$ let $\vec{A}_{j,k} = A_{j,k} e_j \otimes e_k$ be the $K \times K$ matrix with only non-zero entry $A_{j,k}$. This allows us to write

$$D_I A D_I = \textstyle\sum_{i,j} \delta_i \delta_j \vec{A}_{i,j}.$$

Note that for a set I with $|I| = S$ and $A \in \mathbb{R}^{d\times K}$, we have $A_I \in \mathbb{R}^{d\times S}$ and $AD_I \in \mathbb{R}^{d\times K}$. However, as all columns of AD_I that are not in I are zero, the operator norms $\|A_I\|$ and $\|AD_I\|$ coincide.

Chapter 2

Submatrices with non-uniformly selected random supports and insights into sparse approximation

The following chapter essentially is a reprint of the article

S. Ruetz and K. Schnass. Submatrices with nonuniformly selected random supports and insights into sparse approximation. SIAM Journal on Matrix Analysis and Applications, 42(3):1268–1289, 2021

https://doi.org/10.1137/20M1386384

In this chapter we derive concentration inequalities for the operator norms of non-uniformly selected random submatrices. This is especially important in sparse approximation and dictionary learning, which will become apparent in later parts of the thesis where we will make repeated use of the results derived in this chapter. After introducing and discussing some existing results, we provide our new concentration inequalities and show a few applications.

2.1. Introduction

For convenience and motivation, we recall the basic concept of sparse approximation. In sparse approximation, the goal is to find a sparse solution to an underdetermined system of linear equations. A signal $y \in \mathbb{R}^d$ is assumed to be a linear combination of a small number $S \ll d$ of elements ϕ_i, called atoms, out of a larger set, called the dictionary. Denoting the dictionary by $\Phi = (\phi_1, \dots, \phi_K) \in \mathbb{R}^{d \times K}$ and by Φ_I the restriction to the columns indexed by the set I, called the support, one assumes that

$$y \approx \sum_{k \in I} \phi_k x_k = \Phi_I x_I \quad \text{s.t.} \quad |I| = S.$$

The sparse approximation problem amounts to finding the vector x and its support I given the dictionary Φ and signal y. In general, this is a NP-hard problem, hence sparse approxima-

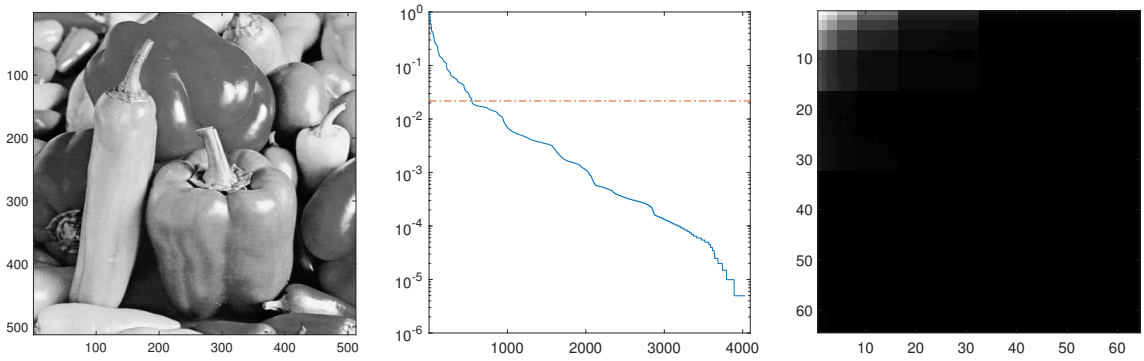

Figure 2.1: Left: Original image from which the patches are extracted. Middle: Relative frequency of wavelet coefficients above threshold (blue) – average frequency (red) on a log scale. Right: Locations of non-zeros coefficients in the 2D Haar-Wavelet basis – the higher the row or column index the smaller the corresponding wavelet.

tion algorithms such as Thresholding, Orthogonal Matching Pursuit (OMP) and Basis Pursuit (BP) were proposed. It turns out that in order to prove support recovery guarantees for these algorithms, information about the extreme singular values of Φ_I is needed.
Let $\|\cdot\|_{2,2}$ denote the operator norm and $\mathbb{I}$ the identity matrix. Deterministic methods to bound $\|\Phi_I^*\Phi_I - \mathbb{I}\|_{2,2}$ for *arbitrary* supports I are of limited use since the restrictions on the dictionary Φ are too stringent. This started the study of random collections of columns of the dictionary Φ. In [82] it was first shown that under rather mild conditions on the dictionary Φ, *most* subdictionaries Φ_I are close to an isometry, i.e., $\|\Phi_I^*\Phi_I - \mathbb{I}\|_{2,2} \leq \vartheta_0 < 1$, with later improvements in [25]. So far, all available results on the conditioning of random subdictionaries rely on the supports I to be drawn from the uniform distribution. Unfortunately this assumption is rarely satisfied for practically relevant signal classes, where some atoms of the underlying dictionary are usually more likely to appear in a sparse representation than others.
To demonstrate this non-homogeneity, we conduct the following small experiment. We take the 2D Haar-Wavelet decomposition of all normalised 64×64 patches from the image *Peppers* and apply a threshold[1] of $\sqrt{\log(d)/(36d)}$ for $d = 64^2$ to the coefficients to get sparse approximations. We then count how often each atom has a non-zero coefficient to get a proxy for its inclusion probability in a sparse support I. 2.1 shows the relative frequency of each element of the 2D Haar-Wavelet basis. It comes as no surprise that low frequency (large) wavelets are much more likely to appear in the sparse supports than high frequency (small) wavelets. So the supports of the sparse signals exhibit a non-uniform structure which previous results on the conditioning of random subdictionaries do not cover. We try to close this gap by defining two non-uniform support distributions and deriving tail bounds on the norms of the resulting random submatrices. This allows us to derive recovery guarantees of the sparse supports for a larger class of practically relevant signals.

Prior work: As mentioned above, Tropp [82] and Chrétien and Darses [25] derived concentration inequalities for the operator norm of random submatrices with uniformly distributed supports. These results were applied to BP showing that BP recovers the correct support and coefficients under rather mild conditions on the dictionary [83]. For OMP, similar results were developed in [73], whereas for Thresholding average case results appeared in [74].
In [48] the dictionary Φ is assumed to be a concatenation of two dictionaries Φ_1 and Φ_2, i.e.,

1. The threshold is inspired by the expected size of the largest inner product of a wavelet with noise drawn uniformly at random from the unit sphere.

$\Phi = (\Phi_1, \Phi_2)$. The authors derive a concentration inequality on the extreme singular values of submatrices that consist of a *fixed* set of columns with cardinality n_1 of the first dictionary and a *random* set of columns n_2 of the second dictionary. This allows to model signals where some atoms are known to be in the support while some others are picked uniformly at random. The idea of using the structure of sparse signals to improve recovery of the sparse coefficients can also be found in the field of compressed sensing (CS). The aim in compressed sensing is to recover a sparse signal $y \in \mathbb{R}^d$ from an incomplete set of linear measurements $z = Ay$, where $A \in \mathbb{R}^{m \times d}$ and $m \ll d$, [18, 30]. The signal y is assumed to be sparse or compressible in some (orthonormal) basis or frame Φ, i.e., $y = \Phi x$ for a sparse coefficient vector x.
From a theoretical point of view the best measurement matrices A, i.e., those achieving the smallest m for a given sparsity level S, are random matrices. Unfortunately in many practical applications it is not possible or efficient to use random matrices, since they cannot be realised by the underlying physical measurement process, such as in compressed magnet resonance imaging (MRI). Instead one is given an (often orthonormal) measurement matrix $\Psi \in \mathbb{R}^{d \times d}$ and has to find a *subsampling pattern* $\Omega \subseteq \{1, ..., d\}$ which selects m rows of Ψ, so that for $A = P_\Omega \Psi$ the signal y and the coefficients x can be reliably reconstructed from $z = Ay = A\Phi x = \bar{A}x$.
As in sparse approximation, rather strong assumptions on the matrix $A\Phi = \bar{A}$ are needed in order to guarantee recovery for all sparse x. In [19] the elements of Ω were assumed to be chosen uniformly at random in order to employ probabilistic arguments to derive sufficient conditions for recovery for relatively small m. Over the years, various different subsampling strategies – most of them highly non-uniform – were proposed (see for example [13, 22, 60, 2, 46, 14]). Underlying the success of these variable density sampling strategies is the highly non-uniform structure of the sparse supports. So it was shown that previous lower bounds on the size of m are too pessimistic and performance can be improved if the subsampling pattern takes the support structure of the sparse signals into account [2, 46, 14].

Contribution: We derive tail bounds for the operator norm of non-uniformly chosen submatrices. The supports are assumed to follow either a Poisson sampling model or a rejective sampling model, thus allowing us to model a large class of non-uniform distributions. Our results rely on a generalisation of a Theorem by Chrétien and Darses [25]. The main tool to handle non-uniformly distributed S-sparse supports is a kind of Poissonisation argument where we provide a generalised version of Lemma 4.1 of [42]. We apply these results to derive sufficient conditions for sparse approximation to work with high probability for Thresholding, OMP and BP. In the CS setup this analysis provides a criterion to decide between two possible measurement matrices A_1 and A_2 depending on the frequency of the basis elements. Further, if there is no design freedom for the dictionary or CS matrix, we show how to incorporate this prior information about the coefficient distribution into the algorithms using the ideas of preconditioning and sensing dictionaries.

Organisation: In Section 2.2 we state our results for norms of non-uniformly distributed random submatrices and apply those concentration inequalities to sparse approximation in Section 2.3. Finally, we incorporate this knowledge in the construction of special sensing dictionaries in Section 2.4 and show how they improve performance.

2.2. Main results

We now present our main results on submatrices whose supports are sampled from a non-uniform distribution. We begin by stating the concentration inequality for the operator norm of non-uniformly picked random submatrices, before turning to some special cases arising in

sparse approximation. Then we state a concentration inequality for the maximal row norm of random column-submatrices. Lastly we state and proof a kind of Poissonisation argument, of independent interest, which is key for our proofs. Note that we state our results only for the rejective sampling model, but they hold for the Poisson sampling model as well – see 2.6.

2.2.1 Operator norm of random submatrices

The aim is to get a tail bound for the random variable $\|H_{I,I}\|_{2,2}$, where I is distributed according to the models introduced above and H is a matrix with zero diagonal. As expected, the result shows how the more frequently picked entries have a higher impact on the operator norm than less important ones.

Theorem 2.1 *Let $H \in \mathbb{C}^{K\times K}$ be a matrix with zero diagonal and assume $I \subseteq \mathbb{K}$ is chosen according to the rejective sampling model with probabilities $p_1, \dots, p_K$ such that $\sum_{i=1}^K p_i = S$. Further let $D_{\sqrt{p}}$ denote the corresponding weight matrix. Then, for all $r \geq 2e^2\|D_{\sqrt{p}}HD_{\sqrt{p}}\|_{2,2}$*

$$\mathbb{P}_S\Big(\|H_{I,I}\|_{2,2} \geq r\Big) \leq 216K \exp\left(-\min\left\{\frac{r^2}{4e^2\|HD_{\sqrt{p}}\|_{\infty,2}^2}, \frac{r^2}{4e^2\|D_{\sqrt{p}}H\|_{2,1}^2}, \frac{r}{2\|H\|_{\infty,1}}\right\}\right).$$

Proof [Proof outline] We follow the proof that appeared in Chrétien and Darses [25] with some minor changes to account for the non-uniformly distributed supports and the extension to non-symmetric matrices. Their proof consists of roughly three steps. First they bound the failure probability of the rejective sampling model by the independent Poisson sampling model, which necessitates Lemma 2.5

$$\mathbb{P}_S\left(\|D_IHD_I\|_{2,2} \geq r\right) \leq 2\mathbb{P}_B\left(\|D_IHD_I\|_{2,2} \geq r\right).$$

Then they use a decoupling argument to make the selection of rows and columns independent, i.e.,

$$\mathbb{P}_B\left(\|D_IHD_I\|_{2,2} \geq r\right) \leq 72\mathbb{P}_B\left(\|D_IHD_I'\|_{2,2} \geq r/2\right),$$

where D_I' is an independent copy of D_I. Then they apply the matrix Chernoff inequality three times to finish the proof. Our proof in the non-uniform, non-symmetric case follows the above outline very closely. The main difficulty lies in bounding the rejective model by the Poisson model, which is why we had to provide Lemma 2.5. The second and third steps are straightforward extensions of their argument. For the sake of completeness we provide a detailed proof in Section 2.5. ■

Special cases – hollow (cross)-Gram matrices

In this subsection we look at the special case $H = \Phi^*\Phi - \mathbb{I}$ that appears naturally in the sparse approximation framework. Previous results showed that success of recovery depends on the coherence $\mu := \max_{i\neq j}|\langle\phi_i, \phi_j\rangle|$ and the conditioning of the subdictionary Φ_I, i.e.,

$$\vartheta_I := \|\Phi_I^*\Phi_I - \mathbb{I}\|_{2,2} = \max\left\{\lambda_{\max}^2(\Phi_I) - 1, 1 - \lambda_{\min}^2(\Phi_I)\right\}.$$

Here $\lambda_{\max}^2(\Phi_I)$ and $\lambda_{\min}^2(\Phi_I)$ denote the largest and smallest eigenvalue of $\Phi_I^*\Phi_I$ respectively. In this setting, the matrix $H := \Phi^*\Phi - \mathbb{I}$ is called the hollow Gram matrix and we call $\mu := \max_{i\neq j}|\langle\phi_i, \phi_j\rangle| = \|H\|_{\infty,1}$ the coherence. Applying Theorem 2.1 to this matrix, we get the following bound on ϑ_I.

Corollary 2.2 *Let $\Phi \in \mathbb{C}^{d\times K}$ be a dictionary with unit norm columns and assume $I \subseteq \mathbb{K}$ is chosen according to the rejective sampling model with probabilities $p_1, \dots, p_K$ such that $\sum_{i=1}^K p_i = S$. Further let $D_{\sqrt{p}}$ denote the corresponding weight matrix. Then, for all $r \geq 2e^2\|D_{\sqrt{p}}HD_{\sqrt{p}}\|_{2,2}$*

$$\mathbb{P}_S\left(\|\Phi_I^*\Phi_I - \mathbb{I}\|_{2,2} \geq r\right) \leq 216K \exp\left(-\min\left\{\frac{r^2}{4e^2\|HD_{\sqrt{p}}\|_{\infty,2}^2}, \frac{r}{2\mu}\right\}\right).$$

In this setting H is symmetric, hence $H^*D_{\sqrt{p}} = HD_{\sqrt{p}}$. The result can be used to bound

$$\mathbb{P}_S\left(\|\Phi_I\|_{2,2} \gtrless \sqrt{1 \pm r}\right) \quad \text{and} \quad \mathbb{P}_S\left(\|(\Phi_I^*\Phi_I)^{-1}\|_{2,2} \geq \frac{1}{1-r}\right).$$

This comes in handy when trying to prove recovery guarantees for sparse approximation algorithms later in this text.
Another frequently arising quantity is the cross-Gram matrix $H := \Psi^*\Phi - \operatorname{diag}(\Psi^*\Phi)$, where Φ and Ψ are dictionaries. In this setting, we call $\hat{\mu} := \max_{i\neq j} |\langle \phi_i, \psi_j\rangle|$ the cross-coherence. Applying Theorem 2.1 yields

Corollary 2.3 *Let $\Psi, \Phi \in \mathbb{C}^{d\times K}$ be dictionaries with unit norm columns and assume $I \subseteq \mathbb{K}$ is chosen according to the rejective sampling model with probabilities $p_1, \dots, p_K$ such that $\sum_{i=1}^K p_i = S$. Further let $D_{\sqrt{p}}$ denote the corresponding weight matrix. Then, for all $r \geq 2e^2\|D_{\sqrt{p}}HD_{\sqrt{p}}\|_{2,2}$*

$$\mathbb{P}_S\left(\|H_{I,I}\|_{2,2} \geq r\right) \leq 216K \exp\left(-\min\left\{\frac{r^2}{4e^2\|HD_{\sqrt{p}}\|_{\infty,2}^2}, \frac{r^2}{4e^2\|D_{\sqrt{p}}H\|_{2,1}^2}, \frac{r}{2\hat{\mu}}\right\}\right).$$

Note that in contrast to Subsection 2.2.1 the matrix H is not symmetric any more, hence we need to control both $\|HD_{\sqrt{p}}\|_{\infty,2}$ and $\|D_{\sqrt{p}}H\|_{2,1}$.
In contrast to previous works the above results are in terms of the maximal row norm of the weighted Gram matrix. Using the bounds

$$\begin{aligned}
&\|HD_{\sqrt{p}}\|_{\infty,2} \leq \|\Psi^*\Phi D_{\sqrt{p}}\|_{\infty,2} \leq \|\Psi^*\|_{\infty,2}\|\Phi D_{\sqrt{p}}\|_{2,2} = \|\Phi D_{\sqrt{p}}\|_{2,2},\\
&\|D_{\sqrt{p}}H\|_{2,1} = \|H^*D_{\sqrt{p}}\|_{\infty,2} \leq \|\Phi^*\|_{\infty,2}\|\Psi D_{\sqrt{p}}\|_{2,2} = \|\Psi D_{\sqrt{p}}\|_{2,2},\\
&\|D_{\sqrt{p}}HD_{\sqrt{p}}\|_{2,2} \leq 2\|\Psi D_{\sqrt{p}}\|_{2,2}\|\Phi D_{\sqrt{p}}\|_{2,2}
\end{aligned}$$

one would get bounds similar in spirit and appearance to [25, 82].
We stick to the quantities $\|HD_{\sqrt{p}}\|_{\infty,2}^2$ and $\|D_{\sqrt{p}}H\|_{2,1}^2$ to see how the weights of the distribution interact with the structure of H. Intuitively the above results state that the more frequently an atom is picked, the less coherent it should be to all the other atoms in order for a random submatrix to be well-conditioned.
The generality of this result allows for $p_i \in [0, 1]$, which thus includes models where some atoms are already known to be in the support and some to not appear at all. This allows for models where a dictionary Φ is a concatenation of two dictionaries Φ_1 and Φ_2, i.e., $\Phi = (\Phi_1, \Phi_2)$ and the submatrix of interest consists of a *fixed* set of columns with cardinality n_1 of the first dictionary and a *random* set of columns n_2 of the second dictionary. Such a scenario can easily be modeled by setting the p_i and the weight matrix $D_{\sqrt{p}}$ accordingly and would yield similar results to [48].

2.2.2 Maximum row norm of a random restriction

Another frequently encountered random variable in sparse approximation is the maximal row norm $\|H_I\|_{\infty,2}$. Given a weight matrix $D_{\sqrt{p}}$, the following Lemma states that one can expect this quantity to be approximately of size $\|HD_{\sqrt{p}}\|_{\infty,2}$. This can be significantly smaller than the worst case $\max_{i,j}|H_{i,j}|\sqrt{S}$ for $|I| \leq S$, depending on the structure of H and $D_{\sqrt{p}}$. Plugging in $H = \Psi^*\Phi - \mathrm{diag}(\Psi^*\Phi)$ we again see that the more frequently picked atoms should have smaller coherences in order for $\|HD_{\sqrt{p}}\|_{\infty,2}$ to be small. This result is an integral part of the proof of Theorem 2.1 and hence we defer its proof to Section 2.5.

Lemma 2.4 *Let $H \in \mathbb{R}^{K\times K}$ be some matrix. Assume $I \subseteq \mathbb{K}$ is chosen according to the rejective sampling model with probabilities $p_1, \ldots, p_K$ such that $\sum_{i=1}^K p_i = S$. Further let $D_{\sqrt{p}}$ denote the corresponding weight matrix. Then, for all $v > 0$*

$$\mathbb{P}_S\left(\|H_I\|_{\infty,2} \geq v\right) \leq 2K\left(e\frac{\|HD_{\sqrt{p}}\|_{\infty,2}^2}{v^2}\right)^{\frac{v^2}{\|H\|_{\infty,1}^2}}.$$

2.2.3 Poissonisation argument in the non-uniform case

As already mentioned, we have to bound the failure probability under the rejective sampling model by the failure probability under the Poisson sampling model in order to apply concentration inequalities for sums of independent random variables. In the uniform case the following lemma is not needed, as one can argue that the supports can also be sampled by drawing one atom after the other to get a uniform support distribution; see Claim (3.29) p. 2173 in [16]. For the non-uniform case it is not that easy. Lemma 4.1 of [42] almost provides the result that we need, but has too restrictive assumptions on the expectations p_i. Therefore we prove[2] the following result which does not have any constraints on the expectations p_i.

Lemma 2.5 (Poissonisation) *Denote by $\mathbb{P}_B$ the probability measure corresponding to the Poisson sampling model* (1.1) *and by $\mathbb{P}_S$ the probability measure corresponding to the rejective sampling model* (1.2) *– both with the same weight matrix $D_{\sqrt{p}}$. Let $f : \mathcal{P}(\mathbb{K}) \mapsto \{0,1\}$ be such that for all $I, J \in \mathcal{P}(\mathbb{K})$*

$$f(I) \leq f(J) \quad \text{if} \quad I \subseteq J.$$

Then for all $I \subseteq \mathbb{K}$ we have $\mathbb{P}_S\left(f(I) = 1\right) \leq 2\,\mathbb{P}_B\left(f(I) = 1\right)$.

Proof Note that the conditions on f imply that if $f(J) = 0$ for some J, then $f(I) = 0$ for all $I \subset J$. We start by showing that for $0 \leq T \leq K-1$ we have

$$\mathbb{P}_B\big(f(I) = 1 \;\big|\; |I| = T\big) \leq \mathbb{P}_B\big(f(I) = 1 \;\big|\; |I| = T+1\big).$$

Expanding the conditional probability we get

$$\frac{\sum_{I:|I|=T} f(I)\mathbb{P}_B(I)}{\sum_{I:|I|=T}\mathbb{P}_B(I)} \leq \frac{\sum_{J:|J|=T+1} f(J)\mathbb{P}_B(J)}{\sum_{J:|J|=T+1}\mathbb{P}_B(J)},$$

which is equivalent to

$$\sum_{I:|I|=T} f(I)\mathbb{P}_B(I) \sum_{J:|J|=T+1}\mathbb{P}_B(J) \leq \sum_{J:|J|=T+1} f(J)\mathbb{P}_B(J) \sum_{I:|I|=T}\mathbb{P}_B(I). \tag{2.1}$$

2. The result might be known but extremely well hidden, thus forcing us to prove it.

Now the crucial step is to see that we can partition these sums in a special way. For a pair (I, J), by definition of the Poisson sampling model, we can write $\mathbb{P}_B(I)\mathbb{P}_B(J)$ in the following way

$$\mathbb{P}_B(I)\mathbb{P}_B(J) = \prod_{i\in I} p_i \prod_{j\notin I}(1-p_j) \prod_{i\in J} p_i \prod_{j\notin J}(1-p_j) = \prod_{i\in I\cap J} p_i^2 \prod_{i\in I\triangle J} p_i(1-p_i) \prod_{j\notin I\cup J}(1-p_j)^2.$$

This implies that for two pairs (I, J), (I', J') with

$$I \cap J = I' \cap J' \quad \text{and} \quad I\triangle J = I'\triangle J' \quad \text{we have} \quad \mathbb{P}_B(I)\mathbb{P}_B(J) = \mathbb{P}_B(I')\mathbb{P}_B(J'),$$

where $I\triangle J$ denotes the symmetric difference between the sets I and J. This allows us to define natural partitions on the set of pairs (I, J) such that the probability $\mathbb{P}_B(I)\mathbb{P}_B(J)$ is constant on each partition: Let $k \in \{0, \dots, T\}$, $A \subseteq \mathbb{K}$ with $|A| = k$ and $B \subseteq \mathbb{K} \setminus A$ with $|B| = 2(T-k)+1$. A will be the intersection and B will model the symmetric difference of the sets I and J respectively. For such a combination of A, B we define

$$\mathcal{Q}_{A,B} := \{(I, J) : I, J \subseteq \mathbb{K}, |I| = T, |J| = T+1, I \cap J = A, I\triangle J = B\}.$$

Note that each pair (I, J) with $|I| = T$, $|J| = T+1$ can be *uniquely* assigned to one $\mathcal{Q}_{A,B}$. So if

$$\sum_{(I,J)\in\mathcal{Q}_{A,B}} f(I) \le \sum_{(I,J)\in\mathcal{Q}_{A,B}} f(J) \tag{2.2}$$

for all possible choices of A, B then (2.1) follows and we are done.
We start with the special case $|A| = 0$ and fix $B \subseteq \mathbb{K}$ with $|B| = 2T+1$. With slight abuse of notation we write $I^c := B \setminus I$ for the complement in B. With this notation (2.2) becomes

$$\sum_{I\subseteq B:|I|=T} f(I) \le \sum_{J\subseteq B:|J|=T+1} f(J).$$

Remembering that $f(I \cup \{i\}) = 1$ if $f(I) = 1$ we get

$$\begin{aligned}
\sum_{I\subseteq B:|I|=T} f(I) &= \sum_{I\subseteq B:|I|=T} f(I)\frac{1}{T+1}\sum_{i\in I^c} f(I) \\
&= \sum_{I\subseteq B:|I|=T} f(I)\frac{1}{T+1}\sum_{i\in I^c} f(I\cup\{i\}) \\
&\le \frac{1}{T+1}\sum_{I\subseteq B:|I|=T}\sum_{i\in I^c} f(I\cup\{i\}) \\
&= \frac{1}{T+1}(T+1)\sum_{J\subseteq B:|J|=T+1} f(J).
\end{aligned}$$

If $|A| > 0$ then the same argument as above replacing $f(\cdot)$ with $f(A \cup \cdot)$ and T with $T - |A|$ yields (2.2) for all possible choices of A and B. Thus we get

$$\mathbb{P}_B\big(f(I) = 1 \;\big|\; |I| = T\big) \le \mathbb{P}_B\big(f(I) = 1 \;\big|\; |I| = T+1\big).$$

Now we are finally in a position to prove our result. Note that

$$\begin{aligned}
\mathbb{P}_B\left(f(I) = 1\right) &= \textstyle\sum_{k=1}^{K} \mathbb{P}_B\left(f(I) = 1 \mid |I| = k\right) \mathbb{P}_B\left(|I| = k\right) \\
&\ge \mathbb{P}_B\left(f(I) = 1 \mid |I| = S\right) \textstyle\sum_{k=S}^{K} \mathbb{P}_B\left(|I| = k\right) \ge \mathbb{P}_S\left(f(I) = 1\right) \cdot \tfrac{1}{2},
\end{aligned}$$

where the last inequality follows from Theorem 3.2 of [45] which says that if the mean number of successes of K independent trials is an integer S, the median is also S. ■

Remark 2.6 *Applying the above result to the functions* $f_1(I) := \mathbb{1}_{\{\|H_{I,I}\|_{2,2} \geq t\}}$ *and* $f_2(I) := \mathbb{1}_{\{\|H_I\|_{\infty,2} \geq t\}}$ *we get*

$$\mathbb{P}_S\left(\|H_{I,I}\|_{2,2} \geq r\right) \leq 2\mathbb{P}_B\left(\|H_{I,I}\|_{2,2} \geq r\right) \qquad \text{and}$$
$$\mathbb{P}_S\left(\|H_I\|_{\infty,2} \geq v\right) \leq 2\mathbb{P}_B\left(\|H_I\|_{\infty,2} \geq v\right).$$

Even though we stated our results only for the rejective sampling model, all of our proofs consist of first bounding the failure probability under the rejective sampling model by the failure probability under the Poisson sampling model. Hence all of our results hold for the Poisson sampling model as well, with the failure bound actually improved by a factor $1/2$.

2.3. Application to sparse approximation

In this section we apply the derived result to sparse approximation. The starting point of sparse approximation is an underdetermined system of linear equations for which one tries to find the sparsest solution. Assuming that the signal y is a linear combination of S columns of a dictionary Φ, we show under which conditions sparse approximation algorithms are successful. To that end we define the following statistical model for our signals.

Definition 2.7 (Signal model) *We model our signals as*

$$y = \Phi_I x_I = \sum_{k=1}^{S} \phi_{i_k} x_{i_k}, \quad x_{i_k} = c_k \sigma_k, \ \ \forall k \in \{1, \dots, S\},$$

where $\Phi \in \mathbb{R}^{d \times K}$ *is a dictionary of* K *normalised atoms,* $I = \{i_1, \dots i_S\}$ *is the random support and* $c = \{c_1, \dots c_S\}$ *is an arbitrary sequence of strictly positive coefficients. We assume* $I \subseteq \mathbb{K}$ *is chosen according to the rejective sampling model with probabilities* $p_1, \dots, p_K$ *such that* $\sum_{i=1}^{K} p_i = S$ *and denote by* $D_{\sqrt{p}}$ *the corresponding weight matrix. Further we assume that the signs* σ_i *form an independent Rademacher sequence, i.e.,* $\sigma_i = \pm 1$ *with equal probability.*

This definition allows us to use probabilistic arguments to show that in the majority of cases, sparse approximation algorithms are able to recover the support under mild conditions on the dictionary Φ and on the coefficients x. We denote by $\mathbb{P}_y := \mathbb{P}_{\sigma,S}$ the product measure of the signs and the support and by $\mu := \max_{i \neq j} |\langle \phi_i, \phi_j \rangle|$ the coherence of the dictionary Φ.

2.3.1 Thresholding

We start by considering the fastest and conceptually easiest sparse approximation algorithm. Thresholding works by finding the indices corresponding to the S largest values of $|\langle y, \phi_i \rangle|$, i.e.,

$$\begin{aligned} &\text{find} \quad J = \operatorname{argmax}_{|I|=S} \|\Phi_I^* y\|_1 \quad \text{and} \\ &\text{reconstruct} \quad x_J = P(\Phi_J) y. \end{aligned}$$

In slight abuse of notation, let $\|c\|_{\min} := \min_i c_i$. In [74], average case results for Thresholding were derived for the uniform case. There, a sufficient condition for Thresholding to work with high probability was $S\mu^2 \log(K) \lesssim \|c\|_{\min}^2 / \|c\|_\infty^2$. We extend these results to the non-uniform case and show how the structure of the dictionary interacts with the distribution of coefficients.

Theorem 2.8 (Thresholding) *Assume that the signals follow the model in* (2.7)*, where the support $I \subseteq \mathbb{K}$ is chosen according to the rejective sampling model with probabilities $p_1, \dots, p_K$ such that $\sum_{i=1}^K p_i = S$. Further let $D_{\sqrt{p}}$ denote the corresponding weight matrix and denote by $H = \Phi^*\Phi - \mathbb{I}$ the hollow Gram-matrix. If*

$$\mu^2 \leq \frac{\|c\|_{\min}^2}{8\|c\|_\infty^2 \log^2(4K/\varepsilon)} \qquad \text{and} \qquad \|HD_{\sqrt{p}}\|_{\infty,2}^2 \leq \frac{\|c\|_{\min}^2}{8e^2\|c\|_\infty^2 \log(4K/\varepsilon)},$$

then Thresholding recovers the support with probability at least $1 - \varepsilon$.

Proof By definition of the algorithm, Thresholding recovers the full support if

$$\|\Phi_{I^c}^* y\|_\infty < \|\Phi_I^* y\|_{\min}.$$

Note that the signals have two sources of randomness, σ and I. Plugging in the definition of y we derive a bound on the failure probability

$$\begin{aligned}
\mathbb{P}_y(\|\Phi_I^* y\|_{\min} < \|\Phi_{I^c}^* y\|_\infty) &= \mathbb{P}_y\left(\|\Phi_I^*\Phi_I x_I\|_{\min} < \|\Phi_{I^c}^*\Phi_I x_I\|_\infty\right) \\
&\leq \mathbb{P}_y\left(\|c\|_{\min} - \|(\Phi_I^*\Phi_I - \mathbb{I})x_I\|_\infty < \|\Phi_{I^c}^*\Phi_I x_I\|_\infty\right) \\
&\leq \mathbb{P}_y\left(\|c\|_{\min} < 2\|H_I x_I\|_\infty\right).
\end{aligned}$$

Where we used that $x_{i_k} = \sigma_k c_k$, where $\sigma \in \mathbb{R}^S$ is an independent Rademacher sequence. Now as the signs σ are independent from the support I, we can apply Hoeffding's inequality to each entry of $H_I\sigma$ (2.19) and use Lemma 2.4 to get

$$\begin{aligned}
\mathbb{P}_y(\|\Phi_I^* y\|_{\min} < \|\Phi_{I^c}^* y\|_\infty) &\leq \mathbb{P}_y\left(\|c\|_{\min} < 2\|H_I x_I\|_\infty\right) \\
&\leq \mathbb{P}_y\left(\|H_I x_I\|_\infty \geq \frac{\|c\|_{\min}}{2} \;\middle|\; \|H_I\|_{\infty,2} < \gamma\right) + \mathbb{P}_S\left(\|H_I\|_{\infty,2} \geq \gamma\right) \\
&\leq 2K \exp\left(-\frac{\|c\|_{\min}^2}{8\|c\|_\infty^2\gamma^2}\right) + 2K\left(e\frac{\|HD_{\sqrt{p}}\|_{\infty,2}^2}{\gamma^2}\right)^{\frac{\gamma^2}{\mu^2}}.
\end{aligned}$$

Setting $\gamma^2 = \frac{\|c\|_{\min}^2}{8\|c\|_\infty^2 \log(4K/\varepsilon)}$, we see that the conditions of the Theorem imply that the failure probability does not exceed ε. ∎

2.3.2 OMP

One of the most popular sparse approximation algorithms is Orthogonal Matching Pursuit (OMP). This greedy algorithm finds the support iteratively, adding one index at a time to the current support. In every step, it picks the index of the atom which has the largest absolute inner product with the residual and then updates the residual. Initialising $r_0 = y$ and $J_0 = \emptyset$, it

$$\begin{aligned}
&\text{finds} \quad j = \operatorname{argmax}_k |\langle \phi_k, r_i \rangle| \quad \text{and} \\
&\text{updates} \quad J_{i+1} = J_i \cup \{j\} \quad \text{resp.} \quad r_{J_{i+1}} = y - P(\Phi_{J_{i+1}})y,
\end{aligned}$$

until a stopping criterion is met. Hence to prove that OMP recovers the correct support, one needs to ensure that it picks an atom from the support in each step. So assuming OMP has successfully found $J \subseteq I$ in the i-th step, it will find another correct atom if

$$\|\Phi_{I^c}^* r_J\|_\infty < \|\Phi_L^* r_J\|_\infty,$$

where $L := I \setminus J$. Based on this observation we prove the following Theorem.

Theorem 2.9 (OMP) *Assume that the signals follow the model in* (2.7)*, where the support* $I \subseteq \mathbb{K}$ *is chosen according to the rejective sampling model with probabilities* $p_1, \dots, p_K$ *such that* $\sum_{i=1}^K p_i = S$*. Let* $D_{\sqrt{p}}$ *denote the corresponding weight matrix. Assume that the hollow Gram-matrix* $H = \Phi^*\Phi - \mathbb{I}$ *satisfies* $\|D_{\sqrt{p}} H D_{\sqrt{p}}\|_{2,2} \leq \frac{1}{4e^2}$*. If*

$$\|HD_{\sqrt{p}}\|_{\infty,2}^2 \leq \min\left\{ \min_{L\subseteq\{1,\dots,S\}} \frac{\|c_L\|_\infty^2}{16e^2\|c_L\|_2^2}, \frac{1}{16e^2\log(216K/\varepsilon)} \right\} \quad \text{and}$$

$$\mu \leq \min\left\{ \min_{L\subseteq\{1,\dots,S\}} \frac{\|c_L\|_\infty}{4\|c_L\|_2\sqrt{\log(218K/\varepsilon)}}, \frac{1}{4\log(218K/\varepsilon)} \right\},$$

then OMP recovers the correct support with probability at least $1-\varepsilon$*.*

Proof Set $\|\Phi_I^*\Phi_I - \mathbb{I}\|_{2,2} =: \vartheta_I$ and assume that $\vartheta_I < 1/2$. We start by expanding the residual in step i

$$r_J = y - P(\Phi_J)y = \Phi_I x_I - P(\Phi_J)\Phi_I x_I = \Phi_{I\setminus J} x_{I\setminus J} - \Phi_J(\Phi_J^*\Phi_J)^{-1}\Phi_J^*\Phi_{I\setminus J}x_{I\setminus J}$$

Set $L := I \setminus J$. By definition, OMP finds another correct atom in the next step if

$$\|\Phi_{I^c}^*(\Phi_L x_L - \Phi_J(\Phi_J^*\Phi_J)^{-1}\Phi_J^*\Phi_L x_L)\|_\infty < \|\Phi_L^*(\Phi_L x_L - \Phi_J(\Phi_J^*\Phi_J)^{-1}\Phi_J^*\Phi_L x_L)\|_\infty,$$

i.e., the inner products with the residual of the remaining atoms in the support are bigger than the inner products with the residual of atoms outside the support. Using the (reverse) triangle inequality, we get the sufficient condition

$$\begin{aligned}\|\Phi_{I^c}^*\Phi_L x_L\|_\infty + \|\Phi_{I^c}^*\Phi_J(\Phi_J^*\Phi_J)^{-1}\Phi_J^*\Phi_L x_L\|_\infty \\ < \|x_L\|_\infty - \|(\Phi_L^*\Phi_L - \mathbb{I})x_L\|_\infty - \|\Phi_L^*\Phi_J(\Phi_J^*\Phi_J)^{-1}\Phi_J^*\Phi_L x_L\|_\infty.\end{aligned}$$

Note that

$$\max\{\|\Phi_{I^c}^*\Phi_L\|_{\infty,2}, \|\Phi_{I^c}^*\Phi_J\|_{\infty,2}, \|\Phi_L^*\Phi_L - \mathbb{I}\|_{\infty,2}, \|\Phi_L^*\Phi_J\|_{\infty,2}\} \leq \|H_I\|_{\infty,2}.$$

So OMP works if

$$2\|H_I\|_{\infty,2}\|x_L\|_2 + 2\|H_I\|_{\infty,2}\|(\Phi_J^*\Phi_J)^{-1}\|_{2,2}\|\Phi_J^*\Phi_L\|_{2,2}\|x_L\|_2 < \|x_L\|_\infty, \tag{2.3}$$

By properties of the operator norm we have $\|\Phi_J^*\Phi_L\|_{2,2} \leq \vartheta_I$ and $\|(\Phi_J^*\Phi_J)^{-1}\|_{2,2} \leq \frac{1}{1-\vartheta_I}$. Plugging this into (2.3) we see that OMP will pick a correct atom in the next step, if

$$\|H_I\|_{\infty,2}\left(2 + 2\frac{\vartheta_I}{1-\vartheta_I}\right) < \frac{\|x_L\|_\infty}{\|x_L\|_2}.$$

So on the set $\{\vartheta_I < 1/2\}$ the columns of Φ_I are linearly independent and we need to have $\|H_I\|_{\infty,2} < \min_{L\subseteq\{1,\dots,S\}} \frac{\|c_L\|_\infty}{4\|c_L\|_2} =: \gamma$ for OMP to find the correct support. So by Corollary 2.2 and Lemma 2.4 we get

$$\begin{aligned}\mathbb{P}_S(\|\Phi_{I^c}^* r_J\|_\infty \geq \|\Phi_L^* r_J\|_\infty) &\leq \mathbb{P}_S(\vartheta_I \geq 1/2) + \mathbb{P}_S(\|H_I\|_{\infty,2} \geq \gamma) \\ &\leq 216K\exp\left(-\min\left\{\frac{1}{16e^2\|HD_{\sqrt{p}}\|_{\infty,2}^2\|}, \frac{1}{4\mu}\right\}\right) + 2K\left(e\frac{\|HD_{\sqrt{p}}\|_{\infty,2}^2}{\gamma^2}\right)^{\frac{\gamma^2}{\mu^2}}.\end{aligned}$$

Owing to the conditions on μ and $\|HD_{\sqrt{p}}\|_{\infty,2}$ in the Theorem, the right hand side does not exceed ε. ∎

Remark 2.10 *Note that for coefficients $c_k \sim \alpha^k$ we can always lower bound $\|c_L\|_\infty / \|c_L\|_2 > \sqrt{1-\alpha^2}$. So in the case of uniformly distributed supports ($p_i = S/K$) and a very incoherent dictionary the conditions above reduce to*

$$S\mu^2 \lesssim 1-\alpha^2 \quad \textit{and} \quad S\mu^2 \log K \lesssim 1,$$

which are essentially the same conditions derived in [73] for exactly sparse signals. This is quite surprising, since this new proof is not only shorter but more importantly does not use the assumption of random signs of the coefficients – meaning that the Theorem also holds without the assumption of random signs.

2.3.3 BP

A very popular alternative to the above algorithms is the Basis Pursuit principle. Instead of tackling the NP-hard problem of finding the sparsest solution with greedy methods, it instead aims to solve the convex relaxation

$$\hat{x} = \operatorname{argmin} \|x\|_1 \quad \text{s.t.} \quad y = \Phi x. \tag{2.4}$$

The average case performance in the uniform case of this optimisation problem has been extensively studied [82, 62, 16]. We give a short proof how these results can be transferred to the non-uniform case.

Theorem 2.11 *Assume that the signals follow the model in (2.7), where the support $I \subseteq \mathbb{K}$ is chosen according to the rejective sampling model with probabilities $p_1, \dots, p_K$ such that $\sum_{i=1}^K p_i = S$. Let $D_{\sqrt{p}}$ denote the corresponding weight matrix. Assume that the hollow Gram-matrix $H = \Phi^*\Phi - \mathbb{I}$ satisfies $\|D_{\sqrt{p}} H D_{\sqrt{p}}\|_{2,2} \leq \frac{1}{4e^2}$. If*

$$\mu \leq \frac{1}{4\log(220K/\varepsilon)} \qquad \textit{and} \qquad \|HD_{\sqrt{p}}\|_{\infty,2}^2 \leq \frac{1}{16e^2 \log(220K/\varepsilon)},$$

then BP recovers the correct coefficients with probability at least $1-\varepsilon$.

Proof We use results for fixed supports such that ℓ_1-minimisation yields the exact solution [81, 36]. In particular we know that if $y = \sum_{i\in I} \phi_i c_i \sigma_i$, for some $I \subset \{1, .., K\}$ with $|I| = S$ and if $\|\Phi_{I^c}^* \Phi_I (\Phi_I^* \Phi_I)^{-1} \sigma_I\|_\infty < 1$, then x is the unique solution to the ℓ_1 - minimisation problem (2.4). So all we have to show is that $\|\Phi_{I^c}^* \Phi_I (\Phi_I^* \Phi_I)^{-1} \sigma_I\|_\infty < 1$ is satisfied with high probability. Set $M := \Phi_{I^c}^* \Phi_I (\Phi_I^* \Phi_I)^{-1}$ and $\vartheta_I := \|\Phi_I^* \Phi_I - \mathbb{I}\|_{2,2}$. As usual we note that

$$\|M\|_{\infty,2} = \|\Phi_{I^c}^* \Phi_I (\Phi_I^* \Phi_I)^{-1}\|_{\infty,2} \leq \|\Phi_{I^c}^* \Phi_I\|_{\infty,2} \|(\Phi_I^* \Phi_I)^{-1}\|_{2,2} \leq \|H_I\|_{\infty,2} \frac{1}{1-\vartheta_I}.$$

Now Corollary 2.2 together with applying Hoeffding's inequality to each entry of $M\sigma$ (2.19) and Lemma 2.4 yields

$$\begin{aligned}
&\mathbb{P}_y\left(\|M\sigma\|_\infty \geq 1\right) \\
&\leq \mathbb{P}_y\left(\|M\sigma\|_\infty \geq 1 \mid \|M\|_{\infty,2} \leq 2\gamma\right) + \mathbb{P}_S\left(\vartheta_I \geq 1/2\right) + \mathbb{P}_S(\|H_I\|_{\infty,2} \geq \gamma) \\
&\leq 2Ke^{-\frac{1}{8\gamma^2}} + 216K \exp\left(-\min\left\{\frac{1}{16e^2\|HD_{\sqrt{p}}\|_{\infty,2}^2}, \frac{1}{4\mu}\right\}\right) + 2K\left(e\frac{\|HD_{\sqrt{p}}\|_{\infty,2}^2}{\gamma^2}\right)^{\frac{\gamma^2}{\mu^2}}.
\end{aligned}$$

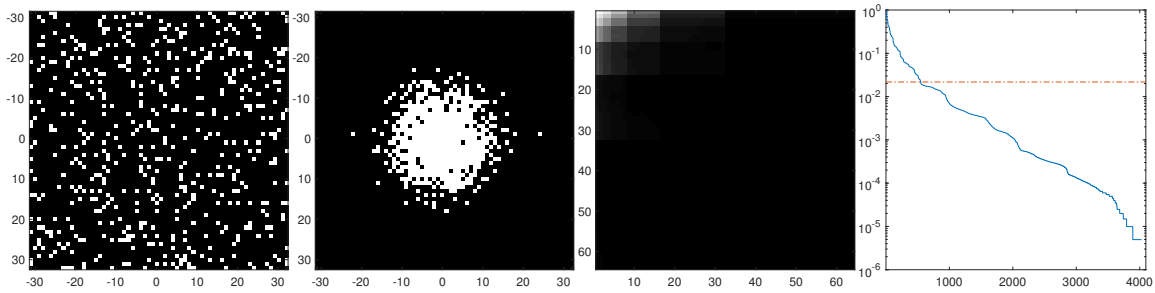

Figure 2.2: From left to right: The K-space $\{(k_1, k_2) : -\sqrt{K}/2+1 \leq k_1, k_2 \leq \sqrt{K}/2\}$ with the frequencies used for the measurement matrix A_1. The frequencies used for the measurement matrix A_2. Locations of non-zero coefficients of patches in the 2D-Haar Wavelet Basis. Expectation of each atom to be in the support (blue) and average expectations for comparison (red) on a log scale.

Setting $\gamma^2 = \frac{1}{8\log(220K/\varepsilon)}$, we see that under the conditions of the Theorem the failure probability is bounded by ε. ■

To illustrate our results we conduct the following small experiment. We take the 2D Haar-Wavelet decomposition of 1000 randomly chosen normalised patches y_n of size 64×64 from the image *Peppers* before applying a threshold of $\sqrt{\log(d)/(36d)}$ for $d = 64^2$ on the coefficients to get sparse approximations. Counting how often each atom is used we get a proxy for the probability of any atom being in the sparse support I, 2.2 (right). We denote by $D_{\sqrt{p}}$ the corresponding weight matrix and by D the vectorised 2D Haar-Wavelet basis. Now we are given two measurement matrices derived from subsampled vectorised 2D-DCT matrices which we denote by $A_1 \in \mathbb{R}^{m\times d}$ and $A_2 \in \mathbb{R}^{m\times d}$. The subsampling pattern is generated by two different subsampling strategies – see 2.2 (left and middle left). For our experiment we set $m = 512$. We are tasked with solving the minimisation problem

$$\hat{x} = \operatorname{argmin} \|x\|_1 \quad \text{s.t.} \quad A_i y = A_i D x \tag{2.5}$$

and are given the choice between the two measurement matrices A_1 and A_2. Our results tell us that as long as the sparse supports of our signals follow the distribution described by the weight matrix $D_{\sqrt{p}}$, we should pick the sensing dictionary A_i that minimises the quantities μ, $\|HD_{\sqrt{p}}\|_{\infty,2}$ and $\|D_{\sqrt{p}}HD_{\sqrt{p}}\|_{2,2}$ (where H is the hollow Gram matrix of the matrix A_iD after normalisation). Looking at 2.1 columns 1-3 we see that for signals following the distribution specified by $D_{\sqrt{p}}$, our results suggest A_2 yields better performance. To test the actual performance, we used BP to recover the coefficients x_n from the set of incomplete measurements $A_i y_n = A_i D x_n$. Note that the coefficients x_n are not sparse, but compressible. Looking at the mean squared error (MSE) $\frac{1}{N}\sum_{n=1}^{N} \|y_n - D\hat{x}_n\|_2^2$ in 2.1, we see that even though strictly speaking our theory does not apply here (as these signals are not perfectly sparse) the quantities $\|HD_{\sqrt{p}}\|_{\infty,2}$ and $\|D_{\sqrt{p}}HD_{\sqrt{p}}\|_{2,2}$ seem to be good predictors of average performance for signals where the sparse support (in this case of the biggest entries) follows a distribution specified by a weight matrix $D_{\sqrt{p}}$.

	μ	$\|HD_{\sqrt{p}}\|_{\infty,2}$	$\|D_{\sqrt{p}}HD_{\sqrt{p}}\|_{2,2}$	MSE
A_1	0.79	3.80	2.80	0.18
A_2	0.98	0.74	0.77	0.06

Table 2.1: The first line corresponds to the uniform subsampling strategy, the second line to the variable density subsampling strategy.

2.4. Sensing dictionaries and preconditioning

As an application of our results we construct a sensing dictionary to improve the average performance of a dictionary for Thresholding and OMP, given that we know the distribution of supports. We then extend these ideas to BP via preconditioning.
In the most general sense a sensing dictionary[3] Ψ for a given dictionary Φ is a matrix of the same size as Φ, whose columns satisfy $\langle \psi_k, \phi_k \rangle = 1$ for all $k \in \mathbb{K}$. It can be used in greedy algorithms to replace the original dictionary in the atom selection step. Sensing dictionaries improving the worst case performance of OMP and Thresholding were first characterised and constructed in [75]. In [74] those ideas were generalised to construct sensing dictionaries that improve the average performance. We extend these average case results to non-uniformly distributed supports to see how the distribution interacts with the structure of the sensing dictionary.
The main idea in Thresholding and OMP is to determine which atoms to include in the support by looking at the absolute inner products between the signal and the atoms. Using a sensing dictionary changes this step in the Thresholding algorithm to

$$\begin{aligned} &\text{find} \quad J = \operatorname{argmax}_{|I|=S} \|\Psi_I^* y\|_1 \quad \text{and} \\ &\text{reconstruct} \quad x_J = P(\Phi_J) y. \end{aligned}$$

For OMP, similarly, the sensing dictionary comes into play when choosing the next atom to add to the support while the residual update step stays the same. Initialising $r_0 = y$ and $J_0 = \emptyset$, for OMP with sensing dictionary Ψ one has to

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

until a stopping criterion is met. Now we will show how to construct a sensing dictionary given knowledge about the distribution of the supports.
Assuming that the distribution of our supports follows a Poisson or rejective sampling model with known weight matrix $D_{\sqrt{p}}$, a sensing dictionary with good average case performance should ideally minimise $\|(\Psi^*\Phi - \mathbb{I})D_{\sqrt{p}}\|_{\infty,2}$ - see Section 2.6. We now try to find Ψ such that this quantity is minimised under the constraint that $\operatorname{diag}(\Psi^*\Phi) = \mathbb{I}$. First note that the quantity $\|(\Psi^*\Phi - \mathbb{I})D_{\sqrt{p}}\|_{\infty,2}^2$ is bounded from above by $\|(\Psi^*\Phi - \mathbb{I})D_{\sqrt{p}}\|_F^2$. Minimising the Frobenius norm instead of the maximum row norm has the big advantage that it is easy to find an analytic solution. For ease of notation let $P := W^2$. Following [74] we use Lagrangian multipliers and derive both the objective and the constraint function along ψ_j to get

$$\frac{d}{d\psi_j}\|\Psi^*\Phi D_{\sqrt{p}}\|_F^2 = \sum_i 2\langle \phi_i, \psi_j \rangle \phi_i p_i = 2\Phi P \Phi^* \psi_j \quad \text{and} \quad \frac{d}{d\psi_j}\langle \phi_j, \psi_j \rangle = \phi_j.$$

3. Note that strictly speaking $\Psi \neq \Phi$ is not actually a dictionary, as the columns are not normalised.

So we see that for Thresholding and OMP, the sensing dictionary should be set to

$$\Psi := (\Phi P \Phi^*)^{-1} \Phi \Lambda,$$

where Λ is a diagonal matrix such that $\langle \phi_i, \psi_i \rangle = 1$ for all $i \in \mathbb{K}$. This compares nicely to the result in [74], where they arrived at $\Psi = (\Phi\Phi^*)^{-1}\Phi\Lambda$ for the special case $p_i = S/K$. This shows how the distribution of coefficients changes the optimal sensing dictionary via the diagonal matrix P. Figures 2.4 and 2.5 show how the performance of Thresholding and OMP improves when using sensing dictionaries for various dictionaries and distributions.
For BP it is not as simple to use a different sensing dictionary. Instead we use preconditioning, multiplying the original dictionary by an invertible matrix from the left and by a diagonal matrix from the right. Inspired by the argument above, we set

$$\Psi = (\Phi P \Phi^*)^{-1/2} \Phi \Lambda,$$

where Λ is a diagonal matrix such that $\langle \psi_i, \psi_i \rangle = 1$ for all $i \in \{1, \dots, K\}$. We then change the BP minimisation problem to

$$\min \|z\|_1 \quad \text{such that} \quad \tilde{y} = \Psi z,$$

where $\tilde{y} = (\Phi P \Phi^*)^{-1/2} y$. This is equivalent to the original optimisation problem, as Λ is a diagonal matrix with positive entries and $(\Phi P \Phi^*)^{-1/2}$ is invertible.

2.4.1 Numerical results

To test the performance of our sensing dictionaries and preconditioning, we conduct the following experiment. We build 2 dictionaries, each with 256 atoms of dimension 128. The columns of the first dictionary are drawn uniformly at random from the unit sphere and the second dictionary is a uniformly subsampled Discrete Cosine Basis with subsequent normalisation. We consider three different distribution models: quadratic, linear and step – see 2.3. For each distribution model and each support size between 1 and 80 we construct 1000 signals by choosing the support according to the rejective sampling model. The sparse coefficients of x

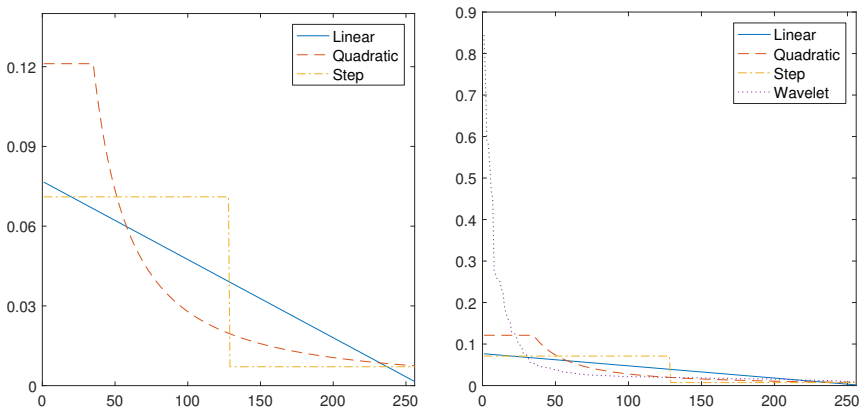

Figure 2.3: Left: Expectations of the Bernoulli random variables employed in our distribution models. Right: The same plot with the relative frequency of the wavelet coefficients from 2.1 for comparison.

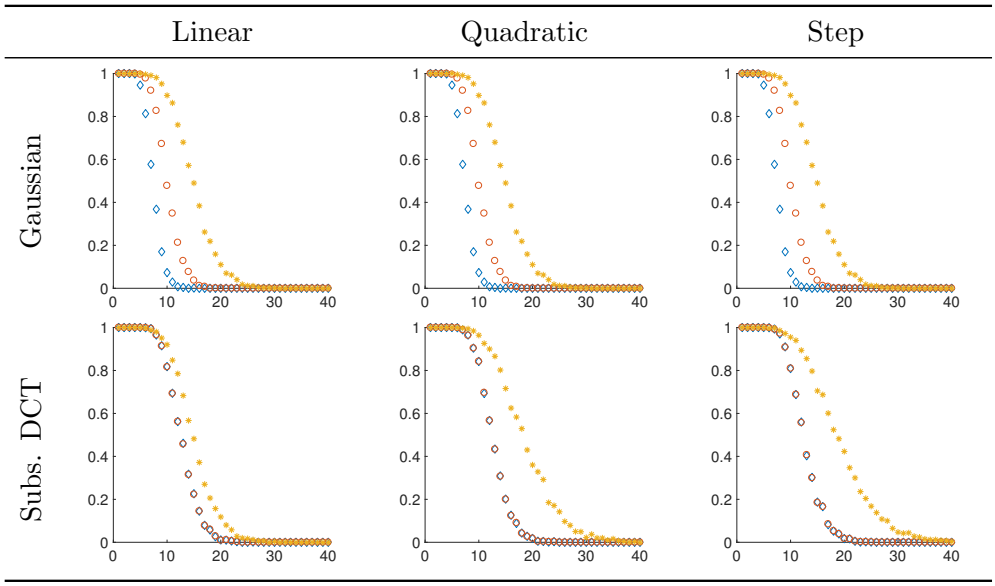

Figure 2.4: Percentage of recovered supports (y-axis) for Thresholding with different sensing dictionaries for various sizes of sparse supports (x-axis). Blue corresponds to no sensing dictionary, red to the uniform average case sensing dictionary and orange to the distribution specific average case sensing dictionary.

have absolute value one with random signs, i.e., $x_i = \pm 1$ with equal probability. We then compare how often Thresholding, OMP and BP can recover the full support when using the original dictionary, the uniform average case sensing dictionary ($P = \mathbb{I}\frac{S}{K}$), and the distribution specific average case sensing dictionary (or the preconditioned matrix for BP). The results for Thresholding and OMP are displayed in 2.4 2.5 respectively. 2.6 shows how the preconditioning changes the recovery rates for BP. As can be seen, incorporating prior knowledge about the distribution of supports into the algorithms improves performance quite significantly for all 3 algorithms.

2.5. Proof of operator norm concentration

The proof follows the one that appeared in Chrétien and Darses [25] with some minor changes to account for the non-uniformly distributed supports and the extension to non-symmetric matrices. We start with an argument that lets us decouple the random variables selecting the rows and columns. This is crucial for the application of concentration inequalities for sums of independent random matrices later in the proof. Throughout this section, we write $\|\cdot\| := \|\cdot\|_{2,2}$ for the operator norm.

Proposition 2.12 *Let $H \in \mathbb{C}^{K\times K}$ be some matrix with zero diagonal. Assume $I \subseteq \mathbb{K}$ is chosen according to the Poisson sampling model with probabilities $p_1, \ldots, p_K$ such that $\sum_{i=1}^{K} p_i = S$. Then, for all $r \geq 0$*

$$\mathbb{P}_B\left(\|D_I H D_I\| \geq r\right) \leq 36\, \mathbb{P}_B\left(\|D_I H D_I'\| \geq r/2\right),$$

where D_I' is an independent copy of D_I.

Proof Let η_i for $1 \leq i \leq K$ be a series of i.i.d. Rademacher random variables. We follow the approach of Chrétien/Darses [25] and Tropp [83] who refer to Bourgain/Tzafriri [12] and de la

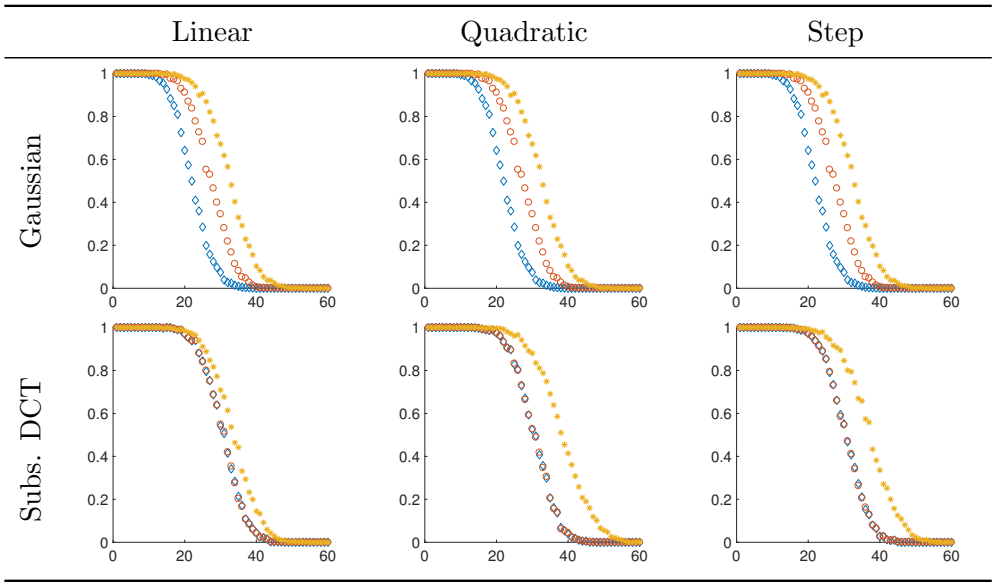

Figure 2.5: Percentage of recovered supports (y-axis) for OMP with different sensing dictionaries for various sizes of sparse supports (x-axis). Blue corresponds to no sensing dictionary, red to the uniform average case sensing dictionary and orange to the distribution specific average case sensing dictionary.

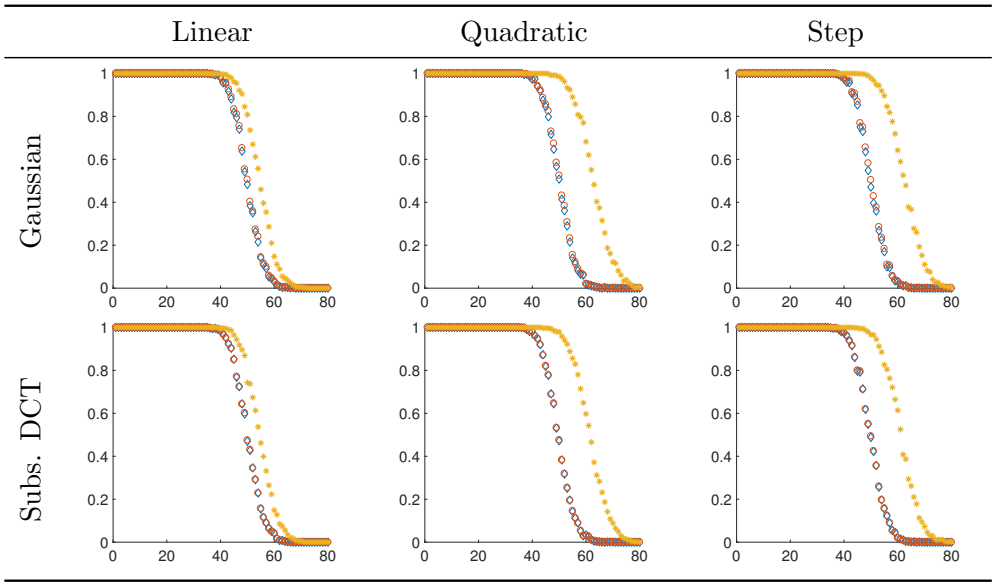

Figure 2.6: Percentage of recovered supports (y-axis) for BP with different preconditioning strategies for various sizes of sparse supports (x-axis). Blue corresponds to the original ℓ_1-minimisation problem, red to preconditioning with uniform weights and orange to preconditioning with the correct weights.

Peña/Giné [28]. We define

$$Z = Z(\eta, \delta) := \sum_{i \neq j} (1 - \eta_i \eta_j) \delta_i \delta_j \vec{H}_{i,j}.$$

Setting $Y = \sum_{i \neq j} \delta_i \delta_j \vec{H}_{i,j} \eta_i \eta_j$, we can write $Z = D_I H D_I - Y$. Recall the Hahn-Banach Theorem.

Theorem 2.13 (Hahn-Banach) *Let X be a real vector space and p a sublinear functional on X. Let f be a linear functional defined on a subspace $A \subset X$ such that $f(a) \leq p(a)$ for all $a \in A$. Then there exists a linear functional $\tilde{f}$ on X satisfying $\tilde{f}(a) = f(a)$ for all $a \in A$ and $\tilde{f}(x) \leq p(x)$ for all $x \in X$.*

From now on we work conditional on a choice of I (i.e. we fix our sequence δ_i, therefore the support set I and the entries of D_I are fixed as well). Denote by $A = \{\lambda D_I H D_I \mid \lambda \in \mathbb{R}\}$ the subspace generated by $D_I H D_I$ and define a linear form $f(\lambda D_I H D_I) = \lambda \|D_I H D_I\|$ on this subspace. By definition we have $f(a) \leq \|a\| =: p(a)$ for all $a \in A$, where the properties of the operator norm imply that p is a sublinear functional. Thus the Hahn-Banach Theorem gives us the existence of a linear functional $\tilde{f}$ satisfying

$$\tilde{f}(D_I H D_I) = f(D_I H D_I) = \|D_I H D_I\| \quad \text{and} \quad \tilde{f}(Z) \leq \|Z\|.$$

Using the linearity of $\tilde{f}$ and that Y is symmetric around 0 we get

$$\begin{aligned}\mathbb{P}_\eta\left(\|Z\| \geq \|D_I H D_I\|\right) &= \mathbb{P}_\eta\left(\|Z\| \geq \tilde{f}(D_I H D_I)\right) \geq \mathbb{P}_\eta\left(\tilde{f}(Z) \geq \tilde{f}(D_I H D_I)\right) \\ &= \mathbb{P}_\eta\left(\tilde{f}(-Y) + \tilde{f}(D_I H D_I) \geq \tilde{f}(D_I H D_I)\right) = \mathbb{P}_\eta\left(\tilde{f}(Y) \geq 0\right),\end{aligned}$$

where again by linearity of $\tilde{f}$ we have

$$\tilde{f}(Y) = \sum_{i \neq j : i,j \in I} \tilde{f}(\vec{H}_{i,j}) \eta_i \eta_j = \sum_{i > j : i,j \in I} \left[\tilde{f}(\vec{H}_{i,j}) + \tilde{f}(\vec{H}_{j,i})\right] \eta_i \eta_j.$$

So we see that $\tilde{f}(Y)$ is a homogeneous Rademacher chaos of order 2. For ease of notation write $\xi := \tilde{f}(Y)$. As ξ is a centered real random variable we can write $\mathbb{E}[|\xi|] = 2\mathbb{E}[\xi \mathbb{I}_{\xi > 0}]$ and a simple application of Hölders inequality yields

$$\mathbb{E}_\eta[|\xi|]^2 = 4\mathbb{E}_\eta[\xi \mathbb{I}_{\xi>0}]^2 \leq 4\mathbb{P}_\eta\left(\xi > 0\right) \mathbb{E}_\eta[\xi^2].$$

Write $\mathbb{E}_\eta[\xi^2] = \mathbb{E}_\eta[\xi^{2/3} \xi^{4/3}]$ and apply Hölders inequality again with $p = \frac{3}{2}$ and $q = 3$ to get

$$\mathbb{E}_\eta[\xi^2] \leq \mathbb{E}_\eta[|\xi|]^{\frac{2}{3}} \mathbb{E}_\eta[\xi^4]^{\frac{1}{3}}.$$

Putting the above together we arrive at

$$\mathbb{P}_\eta(\xi > 0) \geq \frac{1}{4} \frac{\mathbb{E}_\eta[|\xi|]^2}{\mathbb{E}_\eta[\xi^2]} \geq \frac{1}{4} \frac{\mathbb{E}_\eta[\xi^2]^2}{\mathbb{E}_\eta[\xi^4]}.$$

Since ξ is a homogeneous Rademacher chaos of order 2 we can apply Lemma 2.1 of Chrétien and Darses [25], which states

$$\frac{\mathbb{E}_\eta[\xi^2]^2}{\mathbb{E}_\eta[\xi^4]} \geq \frac{1}{9}.$$

So following the arguments in Chrétien and Darses [25] for bounding such a Rademacher chaos we get $\mathbb{P}_\eta\left(\|Z\| \geq \|D_I H D_I\|\right) \geq \frac{1}{36}$. Multiplying both sides with $\mathbb{I}_{\{\|D_I H D_I\| \geq r\}}$ and taking the expectation w.r.t. to I leads to

$$\mathbb{P}_B(\|D_I H D_I\| \geq r) \leq 36\, \mathbb{P}_{B,\eta}(\|Z\| \geq r),$$

where $\mathbb{P}_{B,\eta}$ denotes the product measure. By the same argument as in Proposition 2.1 of [83] there exists a $\bar{\eta} \in \{-1, 1\}^K$ s.t.

$$\mathbb{P}_{B,\eta}\left(\|Z\| \geq r\right) = \mathbb{E}_\eta\left[\mathbb{E}_B\left(\mathbb{I}_{\{\|Z(\eta,\delta)\| \geq r\}} \mid \eta\right)\right] \leq \mathbb{E}_B\left(\mathbb{I}_{\{\|Z(\bar{\eta},\delta)\| \geq r\}}\right) = \mathbb{P}_B\left(\|Z(\bar{\eta},\delta)\| \geq r\right).$$

Setting $T = \{i : \bar{\eta}_i = 1\}$, we see by the definition of Z

$$Z(\bar{\eta}, \delta) = 2 \sum_{j \in T, k \in T^c} \delta_j \delta_k \vec{H}_{j,k} + 2 \sum_{j \in T^c, k \in T} \delta_j \delta_k \vec{H}_{j,k} = 2 \sum_{j \in T, k \in T^c} \delta_j \delta_k (\vec{H}_{j,k} + \vec{H}_{k,j}).$$

Now we can do the decoupling. As δ_i for $i \in T$ are independent from δ_j for $j \in T^c$ we can replace δ_j for $j \in T^c$ with δ' which is an independent copy of δ. Thus

$$\mathbb{P}_{B,\eta}\left(\|Z\| \geq r\right) \leq \mathbb{P}_B\Big(\| \sum_{j \in T, k \in T^c} \delta_j \delta'_k (\vec{H}_{j,k} + \vec{H}_{k,j})\| \geq r/2\Big).$$

The operator norm of this matrix (after reordering) satisfies

$$\left\| \begin{pmatrix} 0 & A \\ B & 0 \end{pmatrix} \right\|^2 = \left\| \begin{pmatrix} B^*B & 0 \\ 0 & A^*A \end{pmatrix} \right\| = \max\{\|A\|^2, \|B\|^2\},$$

where A corresponds to $\sum_{j \in T, k \in T^c} \delta_j \delta'_k \vec{H}_{j,k}$ and B to $\sum_{j \in T, k \in T^c} \delta_j \delta'_k \vec{H}_{k,j}$. As the spectral norm of a submatrix is always less than or equal to the spectral norm of the whole matrix we get by reintroducing the missing entries

$$\mathbb{P}_{B,\eta}(\|Z\| \geq r) \leq \mathbb{P}_B(\|D_I H D'_I\| \geq r/2).$$

Putting everything together yields the desired result. ■

Now we are in a position to apply concentration inequalities for sums of independent random matrices. For that recall the Matrix Chernoff inequality, [84].

Theorem 2.14 (Matrix Chernoff inequality [84]) *Let $X_1, ..., X_K$ be independent, self-adjoint and positive semi-definite random matrices taking values in $\mathbb{C}^{d \times d}$. Now let $B, m > 0$ and assume that for all $1 \leq k \leq K$ $\|X_k\| \leq B$ and $\|\sum_{k=1}^K \mathbb{E} X_k\| \leq m$. Then, for all $t > 0$*

$$\mathbb{P}\left(\|\sum_{k=1}^K X_k\| \geq t\right) \leq d\left(\frac{em}{t}\right)^{t/B}.$$

Now we are going to derive a bound on $\mathbb{P}_B\left(\|D_I H D'_I\| \geq r\right)$ by applying the Matrix Chernoff inequality 3 times. We first use the randomness of D'_I while holding D_I fixed, then we bound the two resulting terms involving D_I. This leads to the following result

Lemma 2.15 *Let $H \in \mathbb{C}^{K \times K}$ be some matrix. Assume $I, I' \subseteq \mathbb{K}$ – leading to the selector matrices R, R' – are chosen according to the Poisson sampling model with probabilities $p_1, \ldots, p_K$ such that $\sum_{i=1}^K p_i = S$. Further let $D_{\sqrt{p}}$ denote the corresponding weight matrix. Then, for all $r > 0$*

$$\mathbb{P}_B\left(\|D_I H D'_I\| \geq r\right) \leq K\left(e\frac{u^2}{r^2}\right)^{\frac{r^2}{v^2}} + K\left(e\frac{\|D_{\sqrt{p}} H D_{\sqrt{p}}\|^2}{u^2}\right)^{\frac{u^2}{\|HD_{\sqrt{p}}\|^2_{\infty,2}}} + K\left(e\frac{\|WH\|^2_{2,1}}{v^2}\right)^{\frac{v^2}{\|H\|^2_{\infty,1}}}.$$

We begin by bounding $\mathbb{P}_B\left(\|D_I H D'_I\| \geq r\right)$.

Lemma 2.16 *Let $H \in \mathbb{C}^{K \times K}$ be some matrix. Assume $I' \subseteq \mathbb{K}$ – leading to the selector matrix R' – is chosen according to the Poisson sampling model with probabilities $p_1, \ldots, p_K$ such that $\sum_{i=1}^K p_i = S$. Further let $D_{\sqrt{p}}$ denote the corresponding weight matrix. Then, for all $r > 0$*

$$\mathbb{P}_B\left(\|D_I H D'_I\| \geq r\right) \leq K\left(e\frac{\|D_I H D_{\sqrt{p}}\|^2}{r^2}\right)^{\frac{r^2}{\|D_I H\|^2_{2,1}}}.$$

Proof Using that for any matrix A we have $\|AA^*\| = \|A^*A\| = \|A\|^2$, we see

$$\mathbb{P}_B\left(\|D_I H D_I'\| > r\right) = \mathbb{P}_B\left(\|D_I H D_I'\|^2 > r^2\right) = \mathbb{P}_B\left(\|D_I H D_I' H^* D_I\| > r^2\right).$$

Denoting by Z_j the j-th column of $D_I H$, we get $D_I H D_I' H^* D_I = \sum_{j=1}^K \delta_j' Z_j Z_j^*$. Then we have $\|Z_j Z_j^*\| = \|Z_j\|_2^2 \leq \|D_I H\|_{2,1}^2$ and

$$\|\sum_{j=1}^K \mathbb{E}[\delta_j' Z_j Z_j^*]\| = \|\sum_{j=1}^K p_j Z_j Z_j^*\| = \|D_I H D_{\sqrt{p}} D_{\sqrt{p}} H^* D_I\| = \|D_I H D_{\sqrt{p}}\|^2.$$

As $\sum_{j=1}^K \delta_j' Z_j Z_j^*$ is a sum of independent random variables, an application of the Matrix Chernoff inequality yields the result. ■

Now we turn to bounding the two quantities $\|D_I H D_{\sqrt{p}}\|$ and $\|D_I H\|_{2,1}$ by the same argument as above.

Lemma 2.17 *Let $H \in \mathbb{R}^{K\times K}$ be some matrix. Assume $I \subseteq \mathbb{K}$ is chosen according to the Poisson sampling model with probabilities $p_1, \ldots, p_K$ such that $\sum_{i=1}^K p_i = S$. Further let $D_{\sqrt{p}}$ denote the corresponding weight matrix. Then, for all $u > 0$*

$$\mathbb{P}_B\left(\|D_I H D_{\sqrt{p}}\| > u\right) \leq K\left(e\frac{\|D_{\sqrt{p}} H D_{\sqrt{p}}\|^2}{u^2}\right)^{\frac{u^2}{\|HD_{\sqrt{p}}\|_{\infty,2}^2}}.$$

Proof Again using that for any matrix A, $\|AA^*\| = \|A^*A\| = \|A\|^2$, we see

$$\mathbb{P}_B\left(\|D_I H D_{\sqrt{p}}\| > u\right) = \mathbb{P}_B\left(\|D_I H D_{\sqrt{p}}\|^2 > u^2\right) = \mathbb{P}_B\left(\|D_{\sqrt{p}} H^* D_I H D_{\sqrt{p}}\| > u^2\right).$$

Now denote by Y_j the j-th row of $HD_{\sqrt{p}}$ then we get $D_{\sqrt{p}} H^* D_I H D_{\sqrt{p}} = \sum_{j=1}^K \delta_j Y_j^* Y_j$. We have $\|Y_j^* Y_j\| = \|Y_j\|_2^2 \leq \|HD_{\sqrt{p}}\|_{\infty,2}^2$ and

$$\|\sum_{j=1}^K \mathbb{E}[\delta_j Y_j^* Y_j]\| = \|\sum_{j=1}^K p_j Y_j^* Y_j\| = \|D_{\sqrt{p}} H^* D_{\sqrt{p}} D_{\sqrt{p}} H D_{\sqrt{p}}\| = \|D_{\sqrt{p}} H D_{\sqrt{p}}\|^2.$$

As $\sum_{j=1}^K \delta_j Y_j^* Y_j$ is a sum of independent random variables, an application of the Matrix Chernoff inequality yields the result. ■

We now restate and prove Lemma 2.4 for the Poisson sampling model. Note that by definition $\|D_I H^*\|_{2,1} = \|HD_I\|_{\infty,2} = \|H_I\|_{\infty,2}$. Recall that by Lemma 2.5

$$\mathbb{P}_S\left(\|H_I\|_{\infty,2} \geq v\right) \leq 2\,\mathbb{P}_B\left(\|H_I\|_{\infty,2} \geq v\right),$$

so this result translates immediately to the rejective sampling model.

Lemma 2.18 *Let $H \in \mathbb{C}^{K\times K}$ be some matrix. Assume $I \subseteq \mathbb{K}$ is chosen according to the Poisson sampling model with probabilities $p_1, \ldots, p_K$ such that $\sum_{i=1}^K p_i = S$. Further let $D_{\sqrt{p}}$ denote the corresponding weight matrix. Then, for all $v > 0$*

$$\mathbb{P}_B\left(\|H_I\|_{\infty,2} \geq v\right) \leq K\left(e\frac{\|HD_{\sqrt{p}}\|_{\infty,2}^2}{v^2}\right)^{\frac{v^2}{\|H\|_{\infty,1}^2}}.$$

Proof We begin by writing $\|H_I\|_{\infty,2}$ as the maximum of a sum of independent random variables $\|H_I\|_{\infty,2}^2 = \max_{i\in\{1,\dots,K\}} \sum_{j=1}^K \delta_j H_{ij}^2$. Now we fix $i \in \{1,..,K\}$ and apply the standard Chernoff inequality

$$\mathbb{P}_B\left(\sum_{j=1}^K \delta_j H_{ji}^2 \geq v^2\right) \leq \left(e\frac{\|HD_{\sqrt{p}}\|_{\infty,2}^2}{v^2}\right)^{\frac{v^2}{\|H\|_{\infty,1}^2}}.$$

Taking a union bound yields the result. ∎

Finally we can put everything together and prove Theorem 2.1. The main difficulty lies in picking v and u such as to minimise the probability bound in 2.15.

Proof [Theorem 2.1] Set

$$\alpha := \min\left\{\frac{r^2}{4e^2\|D_{\sqrt{p}}H\|_{2,1}^2}, \frac{r^2}{4e^2\|HD_{\sqrt{p}}\|_{\infty,2}^2}, \frac{r}{2\mu}\right\} \qquad v^2 := \frac{r^2}{4\alpha} \qquad u^2 := \frac{r^2}{4e^2}.$$

Now these definitions and the assumption $r^2 \geq 4e^4\|D_{\sqrt{p}}HD_{\sqrt{p}}\|^2$ imply

$$\frac{u^2}{\|HD_{\sqrt{p}}\|_{\infty,2}^2} = \frac{r^2}{4e^2\|HD_{\sqrt{p}}\|_{\infty,2}^2} \geq \alpha, \qquad e\frac{\|D_{\sqrt{p}}HD_{\sqrt{p}}\|^2}{u^2} = \frac{4e^3\|D_{\sqrt{p}}HD_{\sqrt{p}}\|^2}{r^2} \leq e^{-1},$$
$$\frac{v^2}{\mu^2} = \frac{r^2}{4\alpha\mu^2} \geq \alpha, \qquad e\frac{\|D_{\sqrt{p}}H\|_{2,1}^2}{v^2} = \frac{4e\|D_{\sqrt{p}}H\|_{2,1}^2\alpha}{r^2} \leq e^{-1},$$
$$\frac{r^2}{4v^2} = \frac{4r^2\alpha}{4r^2} = \alpha, \qquad e\frac{4u^2}{r^2} = \frac{4er^2}{4e^2r^2} = e^{-1}.$$

So $\mathbb{P}_S\left(\|D_IHD_I\| \geq r\right) \leq 2\mathbb{P}_B\left(\|D_IHD_I\| \geq r\right) \leq 72\mathbb{P}_B\left(\|D_IHD_I'\| \geq r/2\right)$, together with

$$\mathbb{P}_B\left(\|D_IHD_I'\| \geq r\right) \leq K\left(\left(e\frac{4u^2}{r^2}\right)^{\frac{r^2}{4v^2}} + \left(e\frac{\|D_{\sqrt{p}}HD_{\sqrt{p}}\|^2}{u^2}\right)^{\frac{u^2}{\|HD_{\sqrt{p}}\|_{\infty,2}^2}} + \left(e\frac{\|D_{\sqrt{p}}H\|_{2,1}^2}{v^2}\right)^{\frac{v^2}{\mu^2}}\right)$$

shows that $\mathbb{P}_S\left(\|D_IHD_I\| \geq r\right) \leq 216Ke^{-\alpha}$. ∎

For convenience we restate an easy consequence of Hoeffding's inequality.

Lemma 2.19 (Hoeffding) *Let $M \in \mathbb{R}^{K\times S}$ be a matrix and $x \in \mathbb{R}^S$ such that $\operatorname{sign}(x) \in \mathbb{R}^S$ is an independent Rademacher sequence. Then, for all $t \geq 0$*

$$\mathbb{P}_\sigma\left(\|Mx\|_\infty \geq t\right) \leq 2K\exp\left(-\frac{t^2}{2\|M\|_{\infty,2}^2\|x\|_\infty^2}\right).$$

Proof We apply Hoeffding's inequality to the k-th entry of Mx, which yields

$$\mathbb{P}_\sigma\left(|\sum_j M_{k,j}x_j| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\sum_j M_{k,j}^2x_j^2}\right) \leq 2\exp\left(-\frac{t^2}{2\|x\|_\infty^2\|M_{:,k}\|_2^2}\right).$$

The statement follows using a union bound and the identity $\|M\|_{\infty,2} = \max_k \|M_{:,k}\|_2$. ∎

Remark 2.20 *In the published version of Chrétien and Darses [25] there is a tiny bug in the proof of Proposition 4.2 in the way the variables u and v are balanced. In particular, for very small μ, inequality 4.17 may be violated. v^2 should instead be defined via an equality in 4.15, whereas 4.14 should be an inequality.*

2.6. Sensing matrices

Lemma 2.21 (Thresholding with sensing matrix) *Assume that the signals follow the model in* (2.7)*, where the support $I \subseteq \mathbb{K}$ is chosen according to the rejective sampling model with probabilities $p_1, \dots, p_K$ such that $\sum_{i=1}^K p_i = S$. Further let $D_{\sqrt{p}}$ denote the corresponding weight matrix and denote by $H := \Psi^*\Phi - \mathbb{I}$ the hollow cross-Gram matrix. If*

$$\|H\|_{\infty,1}^2 \leq \frac{\|c\|_{\min}^2}{8\|c\|_{\max}^2 \log(4K/\varepsilon)}, \qquad \text{and} \qquad \|HD_{\sqrt{p}}\|_{\infty,2}^2 \leq \frac{\|c\|_{\min}^2}{8e^2\|c\|_{\max}^2 \log(4K/\varepsilon)},$$

then Thresholding with sensing dictionary Ψ recovers the support with probability at least $1-\varepsilon$.

Proof Now by definition of the algorithm, Thresholding recovers the full support if

$$\|\Psi_{I^c}^* y\|_{\max} < \|\Psi_I^* y\|_{\min}.$$

Repeating the steps from the proof of Theorem 2.8 with the obvious changes we obtain the result. ∎

Lemma 2.22 (OMP with sensing matrix) *Assume that the signals follow the model in* (2.7)*, where the support $I \subseteq \mathbb{K}$ is chosen according to the rejective sampling model with probabilities $p_1, \dots, p_K$ such that $\sum_{i=1}^K p_i = S$. Further let $D_{\sqrt{p}}$ denote the corresponding weight matrix. Let Ψ be a sensing matrix and assume the hollow Gram-matrix $H = \Phi^*\Phi - \mathbb{I}$ satisfies $\|D_{\sqrt{p}} H D_{\sqrt{p}}\|_{2,2} \leq \frac{1}{4e^2}$. If*

$$\begin{aligned}
\|HD_{\sqrt{p}}\|_{\infty,2}^2 &\leq \frac{1}{16e^2 \log(216K/\varepsilon)} \\
\|H\|_{\infty,1} &\leq \frac{1}{4\log(218K/\varepsilon)} \\
\|(\Psi^*\Phi - \mathbb{I})D_{\sqrt{p}}\|_{\infty,2}^2 &\leq \min_{L \subseteq \{1,\dots,S\}} \frac{\|c_L\|_\infty^2}{16e^2\|c_L\|_2^2} \\
\|\Psi^*\Phi - \mathbb{I}\|_{\infty,1} &\leq \min_{L \subseteq \{1,\dots,S\}} \frac{\|c_L\|_\infty}{4\|c_L\|_2\sqrt{\log(218K/\varepsilon)}},
\end{aligned}$$

then OMP with sensing matrix Ψ recovers the correct support with probability at least $1-\varepsilon$.

Proof Set $L := I \setminus J$. By definition, OMP finds another correct atom in the next step if

$$\|\Psi_{I^c}^*(\Phi_L x_L - \Phi_J(\Phi_J^*\Phi_J)^{-1}\Phi_J^*\Phi_L x_L)\|_\infty < \|\Psi_L^*(\Phi_L x_L - \Phi_J(\Phi_J^*\Phi_J)^{-1}\Phi_J^*\Phi_L x_L)\|_\infty.$$

Repeating the steps from the proof of Theorem 2.9 with the obvious changes we obtain the result. ∎

2.7. Discussion

In this chapter we have derived concentration inequalities for norms of random subdictionaries with non-uniformly distributed sparse supports. This has allowed us to derive sufficient conditions for sparse approximation algorithms to recover the correct support with high probability

given that the supports follow a rejective sampling or Poisson sampling model. We have shown that recovery of signals depends on the structure of the cross-Gram matrix and the distribution of supports, proving that more frequently used atoms should be more incoherent than less frequently used ones. The generalisation from uniformly to non-uniformly distributed supports gives valuable insight into how, in a compressed sensing setup, measurement matrices should be chosen or constructed. For both Thresholding and OMP it was shown that using sensing dictionaries that take the distribution of supports into account improves performance. Using preconditioning to extend this argument to BP, it was shown that prior knowledge about the distribution also leads to improved performance for BP. In the next chapter we will see how the results of this chapter can be applied in practice.

Chapter 3

Adapted variable density subsampling for compressed sensing

Equipped with the knowledge of how to bound operator norms of non-uniformly selected random submatrices we turn to a prime application — compressed sensing. We will show how to derive optimal subsampling strategies in a variable density setup by characterising the sparsity patterns of our signals via a (possibly non-uniform) distribution on their supports. The results in this chapter are based on [66].

3.1. Compressed Sensing

Let $x \in \mathbb{C}^K$ be some signal and $A \in \mathbb{C}^{m \times K}$ be some matrix, usually called the measurement matrix. Compressed sensing (CS) consists of reconstructing the signal x from measurements $y = Ax$. Usually $m < K$ and it is assumed that the signal x is S-sparse, meaning that only $S \ll K$ elements of x are non-zero. One tries to recover x by solving the following optimisation problem

$$\hat{x} = \arg\min \|x\|_1 \quad \text{s.t.} \quad y = Ax. \tag{3.1}$$

Starting with the seminal works [17, 32], compressed sensing theory tries to find sufficient conditions for the above minimisation problem to recover the sparse signal. Early results suggested that if each entry of the matrix A is sampled i.i.d. from a Gaussian distribution and $m \gtrsim S \log(K)$, then the above minimisation program does yield the correct solution with high probability.
These results were very soon extended to a random subsampling setting, where the sensing matrix A is constructed by sampling rows a_k from a unitary matrix $A_0 \in \mathbb{C}^{K \times K}$ uniformly at random [19, 63]. In this setting, a typical sufficient condition for the above minimisation problem to recover the sparse signal with probability at least $1 - \varepsilon$ reads as

$$m \gtrsim SK \max_{1 \leq k \leq K} \|a_k\|_\infty^2 \log(K/\varepsilon). \tag{3.2}$$

If A_0 is the discrete Fourier matrix — for which $\max_{1 \leq k \leq K} \|a_k\|_\infty^2 = \frac{1}{K}$ — this leads to theoretical results comparable to the Gaussian setting. Nevertheless this still falls short of explaining the remarkable success of CS in most applications, as $K \max_{1 \leq k \leq K} \|a_k\|_\infty^2$ is usually quite large.
To solve this problem, variable density subsampling was introduced [63, 23, 60, 22, 46]. There

the sensing matrix $A \in \mathbb{C}^{m\times K}$ is constructed by sampling the rows of A_0 via a (possibly non-uniform) probability distribution. Concretely, the sensing matrix A is defined to be

$$A := \frac{1}{\sqrt{m}} \left(\frac{1}{\sqrt{\omega_{j_\ell}}} a_{j_\ell} \right)_{1\leq \ell \leq m},$$

where m is the number of measurements we are allowed to take and j_ℓ for $1 \leq \ell \leq m$ are i.i.d random variables such that $\mathbb{P}(j_\ell = k) = \omega_k$. Note that the subsampling strategy is determined by the probabilities ω_k for $1 \leq k \leq K$. A typical choice in this setting is $\omega_k := \frac{\|a_k\|_\infty^2}{\sum_k \|a_k\|_\infty^2}$ leading to the sufficient condition

$$m \gtrsim S \sum_k \|a_k\|_\infty^2 \log(K/\varepsilon).$$

However, this still does not completely bridge the gap between theory and application. Recent results go further by arguing that the optimal subsampling strategy should not only depend on the sensing and sparsity matrices, but also on the structure of the sparse signals [2, 14, 3]. The so called flip test proposed in [2] beautifully illustrates this. The assumption of knowledge of the structure of the sparse signals was shown to be especially important in the case of blocks of measurements [14, 26, 3]. The drawback of all of these results is that they rely on the exact knowledge of the locations of the non-zero coefficients of the sparse signal, which by definition of the problem is not available in practice.

3.2. Contribution

We generalise the aforementioned results and show that the subsampling strategy should depend on the structure of the sensing/sparsity matrix together with the **distribution** of sparse supports. In practice, if one has access to a number of signals from the same signal class as x, a guess of the underlying distribution of sparse supports of x can be made and the optimal subsampling pattern be thus derived. We then extend our results to the setting of structured acquisition, where instead of isolated measurements, blocks of measurements are taken. In Section 3.3 the main result is stated, Section 3.4 applies our theory to some special cases to compare it to existing results and Section 3.5 shows how to apply the theory in practice. Section 3.6.2 looks at the setting of sparsity in levels and blocks of measurements in more detail. The proof of our main result is stated in Section 3.7.

3.3. Main result

Assume we are given a unitary matrix $A_0 \in \mathbb{C}^{K\times K}$ representing the set of possible linear measurements $(A_0^*)_i =: a_i^*$. We partition the set $\mathbb{K}$ into M blocks $\mathcal{I}_k$ such that $\uplus_k \mathcal{I}_k = \mathbb{K}$ and set

$$B_k := (a_i)_{i\in\mathcal{I}_k} \in \mathbb{C}^{|\mathcal{I}_k|\times K}$$

The sensing matrix A is then defined as

$$A := \frac{1}{\sqrt{m}} \left(\frac{1}{\sqrt{\omega_{j_\ell}}} B_{j_\ell} \right)_{1\leq \ell \leq m},$$

where $m \leq M$ is the number of blocks we want to measure and j_ℓ for $1 \leq \ell \leq m$ are i.i.d random variables such that $\mathbb{P}(j_\ell = k) = \omega_k$. So the ω_k define the probability with which each

block of measurements is selected. In line with existing compressed sensing literature we call

$$\max_k \|a_k\|_\infty^2, \tag{3.3}$$

the coherence of the matrix A_0. This is not the usual coherence used in other chapters of this thesis, but can be seen as cross-coherence $\mu(A_0^*, \mathbb{I}) = \max_{k,j} |\langle a_k, e_j \rangle|$.

Definition 3.1 (Signal model) *We model our signals as*

$$x = \sum_{i \in I} e_i x_i \sigma_i, \tag{3.4}$$

where $x_i \in \mathbb{R}$ (or $\mathbb{C}$) and $I = \{i_1, \dots i_S\}$ is the random support following either the rejective or Poisson sampling model with weight vector ω such that $\sum_{i=1}^K \omega_i = S$ and denote by D_ω the corresponding diagonal matrix. Further we assume that σ_i forms a Rademacher (or Steinhaus[1]) sequence.

With these definitions we are finally able to state our main result.

Theorem 3.2 *Assume that the signals follow the model in 3.1, where the support $I \subseteq \mathbb{K}$ is chosen according to the rejective sampling model with probabilities $p_1, \cdots, p_K$ such that $\sum_{k=1}^K p_k = S$ and $0 < p_k \leq 1$. If the measurements B_k are sampled according to probabilities ω_k and if*

$$\begin{aligned} m &\gtrsim \max_k \frac{\|B_k^* B_k\|_{\infty,1}}{\omega_k} \log^3(K/\varepsilon), \\ m &\gtrsim \max_k \frac{\|B_k D_p B_k^*\|_{2,2}}{\omega_k} \log^2(K/\varepsilon), \end{aligned} \tag{3.5}$$

then (3.1) *recovers the sparse signal with probability $1 - \varepsilon$.*

The exact statement — including constants — can be found in Section 3.7. The restriction $p > 0$ is no real constraint, as in the case of $p_i = 0$ for some i, a careful analysis of the proof shows that one can then set the columns of A with indices in J^c to zero since they are never part of the random supports I anyway.

Remark 3.3 *This result also extends to signals x that are sparse in some unitary basis Ψ by change of variable. If we denote by Φ^* our original sensing matrix and let $x = \Psi z$ for some sparse vector z, then we can again apply the above result with the new sensing matrix $A_0 = \Phi^* \Psi$ and sparse signal z. In this case, the coherence $\|a_k\|_\infty$ really is similar to a cross-coherence by noting that $\|a_k\|_\infty = \max_{i,j} |\langle \phi_i, \psi_j \rangle|$.*

The above result shows that the optimal sampling strategy ω should depend both on the distribution p of sparse supports via the diagonal matrix D_p and on the structure of the blocks B_k. One way to optimise the above bounds is by setting

$$\omega_k := \frac{\max\{\|B_k D_p B_k^*\|_{2,2}, \|B_k^* B_k\|_{\infty,1}\}}{L}, \tag{3.6}$$

1. Meaning σ_i are independent realisations of the random variable e^{iZ} with Z uniformly distributed on $(0, 2\pi)$

where L is a normalising constant ensuring $\sum_k \omega_k = 1$. By plugging this bound into the above theorem we get that we need about

$$m \gtrsim \left(\sum_k \|B_k D_p B_k^*\|_{2,2} + \sum_k \|B_k^* B_k\|_{\infty,1} \right) \log^3(K/\varepsilon) \tag{3.7}$$

measurements to ensure recovery with high probability. In Section 3.4 we will look at special cases of blocks of measurements, where this bound on m can be further simplified. For isolated measurements, i.e. $B_k = a_k$ the above can be further simplified to yield the following result.

Corollary 3.4 *Assume that the signals follow the model in 3.1, where the support $I \subseteq \mathbb{K}$ is chosen according to the rejective sampling model with probabilities $p_1, \cdots, p_K$ such that $\sum_{k=1}^K p_k = S$ and $0 < p_k \leq 1$. If the measurements a_k are sampled according to*

$$\omega_k = \frac{\max\{a_k D_p a_k^*, \|a_k\|_\infty^2\}}{L}, \tag{3.8}$$

where L is a normalising constant ensuring $\sum_k \omega_k = 1$, and if

$$m \gtrsim \left(S + \sum_k \|a_k\|_\infty^2 \right) \log^3(K/\varepsilon), \tag{3.9}$$

then (3.1) *recovers the sparse signal with probability $1 - \varepsilon$.*

Proof First note that $\|B_k D_p B_k^*\|_{2,2} = a_k D_p a_k^*$ and thus

$$\sum_k a_k D_p a_k^* = \operatorname{tr}(A_0 D_p A_0^*) = \operatorname{tr}(D_p) = S.$$

Further

$$\|B_k^* B_k\|_{\infty,1} = \|a_k^* a_k\|_{\infty,1} \leq \max_{i,j} |a_{k,i} a_{k,j}| \leq \max_i |a_{k,i}|^2 = \|a_k\|_\infty^2,$$

leading to $L \leq S + \sum_k \|a_k\|_\infty^2$. Plugging these ω_k into Theorem 3.2 yields the result. ∎

This result is an improvement upon standard results for general (unknown) supports I, which read as $m \gtrsim S \sum_k \|a_k\|_\infty^2 \log(K)$ [19, 60, 46, 22]. This is to be expected since we assume that information about the supports and their distribution is available. On the other hand, the additional log factors are the price we pay for our random signal approach. A comparison to existing results that assume knowledge about the structure of sparsity, which will be done in the next section, will thus be more interesting.
Further, Corollary 3.4 shows how, for a given weight vector p, this lower bound is attained via the formula in (3.8). This is an easy-to-use recipe yielding state of the art results in a number of experiments (see Sections 3.4 and 3.5). Before moving on to empirical results, we want to mention a few special cases of measurement matrices, sparsity basis and weights p which underline the generality of the above result.

3.4. Special cases

In this section we show how our result can be applied to recover state of the art theoretical results in CS theory.

3.4.1 Coherent matrix

A frequent example showing the necessity of some sort of knowledge of the structure in sparse signals is the special case where $A_0 = \mathbb{I}$. Denote by $J := \{i : p_i \neq 0\}$ the set of indices, where the weights of our random support model are zero and set the columns of A_{J^c} to zero. In this setting, Formula (3.8) leads to $\omega_k = \frac{\delta_{k,J}}{|J|}$ and thus $m \gtrsim |J| \log^3(K/\varepsilon)$ which means that to ensure recovery with high probability, we have to sample all rows corresponding to positive weights p_ℓ, i.e. all those rows that correspond to entries of our sparse vector that have a non-zero probability of appearing in the support. This also includes the setting where $p \in \{0, 1\}$ recovering, up to logarithmic factors, results derived in [14] for fixed sparse supports.

3.4.2 Fourier matrix

Assume that $A_0 = \mathcal{F}$, i.e. the 1-D Fourier transform. This matrix is known to be incoherent ($\|a_k\|_\infty^2 = \frac{1}{K}$) and in the isolated measurement setting this yields $a_k D_p a_k^* = \sum_\ell |a_{k,\ell}| p_\ell \leq \|a_k\|_\infty \|p\|_1 = \frac{1}{K} \sum_\ell p_\ell = \frac{S}{K}$ for any weight vector p (recall that we have $\sum_\ell p_\ell = S$). Plugging these observations back into our main Theorem yields that independently of the distribution p, one should sample uniformly at random, i.e. $\omega_k = \frac{1}{K}$. Corollary 3.4 thus yields $m \gtrsim S \log^3(K)$ which (up to log factors) is in line with standard lower bounds on the number of measurements [17, 31].

3.4.3 Uniformly distributed sparse supports

One possible distribution of our sparse supports is the uniform distribution, where $p_\ell = S/K$. Plugging this into Formula (3.8) yields

$$\omega_k = \frac{\max\{S/K, \|a_k\|_\infty^2\}}{L},$$

where L again is a normalising constant. This is very similar in spirit to coherence based subsampling strategies [63, 23, 60], where $\omega_k := \frac{\|a_k\|_\infty^2}{\sum_\ell \|a_\ell\|_\infty^2}$. Since in the uniform case there is no structure in the sparse signals that can influence the subsampling strategy it is only natural that in this special case the optimal subsampling strategy depends only on the structure of the sensing matrix together with a lower bound S/K to cover the whole space.
We conduct a small experiment by setting $K = 2^{16}$ and $S = \sqrt{K}/2$. Further we let Φ be the 2D Hadamard transform and Ψ be the 2D Haar wavelet transform. We then generate 100 synthetic signals with uniformly distributed sparse supports, coefficients with absolute value 1 and random signs to compare the performance of three different subsampling strategies, which can be seen in Figure 3.1. Sampling 5% of measurements from each of these distributions and subsequently solving 3.1 with the Nesta algorithm [55, 10] and averaging the PSNR over 10 runs, each with 100 fresh signals, shows that our adapted subsampling strategy outperforms both the uniform and the coherence bases subsampling strategy. This shows that in this special case our result is tight in the sense that both terms in the numerator of Formula 3.8 are indeed necessary.

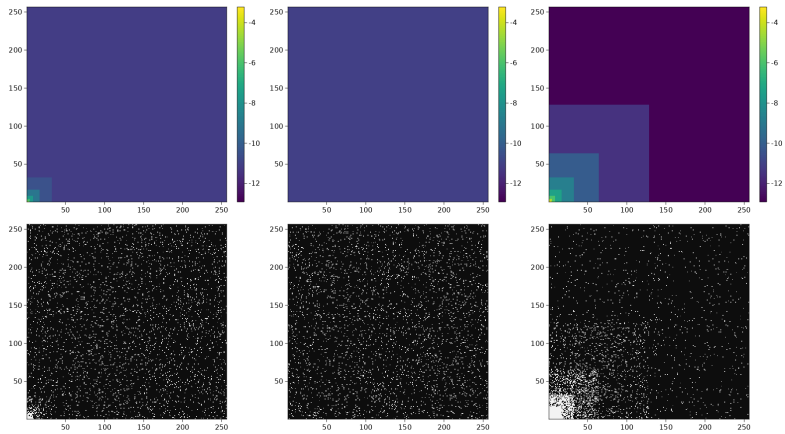

Figure 3.1: Subsampling densities (top row) and corresponding samples (bottom row) for the adapted variable density sampling scheme (left column), the uniform distribution (middle right) and the coherence based subsampling scheme (right row). The resulting average PSNR are: Adapted - 133.5, Uniform - 105.6 and Coherence - 62.3.

3.5. Numerical experiments

Now that we conduct a few experiments, in each of which we assume to be given a training set of images from which we generate the sparse distribution model by transforming them into a wavelet basis before applying a threshold. The relative frequency with which each coefficient appears in these sparse supports is our proxy for the inclusion probabilities p, since they are just the expectation of an atom being in the support. This one-to-one correspondence is also motivated by the close relationship between the rejective sampling model and the Bernoulli sampling model with weights p. We further assume to be given a reference image which we have to reconstruct. We will compare the performance of our subsampling strategy in the isolated measurement case against a state-of-the-art variable density subsampling scheme with polynomial decay, where we pick a frequency (k_1, k_2) in the 2D k-space with probability $\frac{1}{(k_1^2+k_2^2)^{2.5}}$. To ensure meaningful results, each experiment is averaged over 10 runs. We will use the 2D Fourier matrix to take measurements and plot all sampling distributions in log-scale.
For our first experiment (Figure 3.2) we assume a standard compressed sensing setup with isolated 2D Fourier measurements and a 2D DB4 wavelet matrix as sparsifying basis. We want to sense the reference brain image (bottom right). To approximate the distribution of the sparse supports, we use a dataset of around 4.000 real brain images [15] onto which we apply the 2D DB4 wavelet transform followed by a thresholding operation with a threshold of around 0.006, yielding the matrix W (top right). Plugging these weights into Formula (3.8) and normalising the resulting density to 1, we get the adapted subsampling distribution ω (top left). We compare this strategy to the above mentioned polynomial decaying density (top middle) by sampling 10% of frequencies in the k-space (bottom left and middle). Finally, an application of the Nesta algorithm to solve (3.1) for both sets of measurements yields the results in the figure. As can be seen, the adapted subsampling strategy is able to slightly outperform the quadratically decaying subsampling strategy — resulting in a PSNR value of 23.8 compared to 32.0.
To show that our new subsampling strategy does indeed adapt to the underlying distribution of sparse supports, we repeat the above experiment (Figure 3.3) but this time use a different dataset — the MRNet dataset which consists of around 30.000 images of knees [41]. To generate

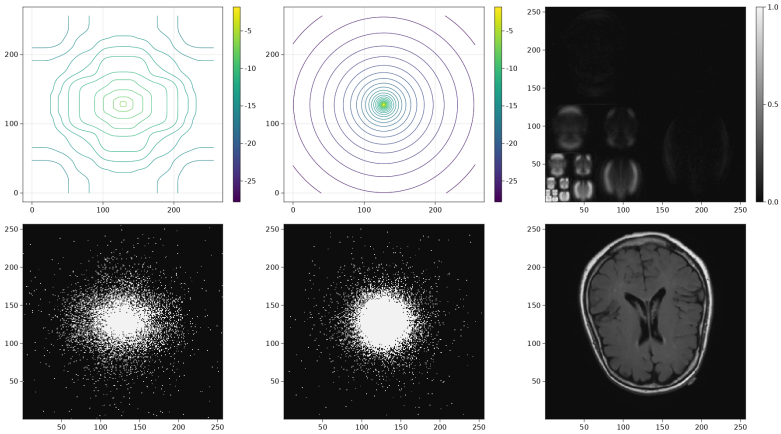

Figure 3.2: Adapted variable density sampling scheme (left column) vs polynomial decay (middle column). Matrix W of sparse support distribution in the DB4 wavelet basis (top right) and test image (bottom right). The resulting PSNR values are: Adapted - 32.8 and Polynomial - 32.0.

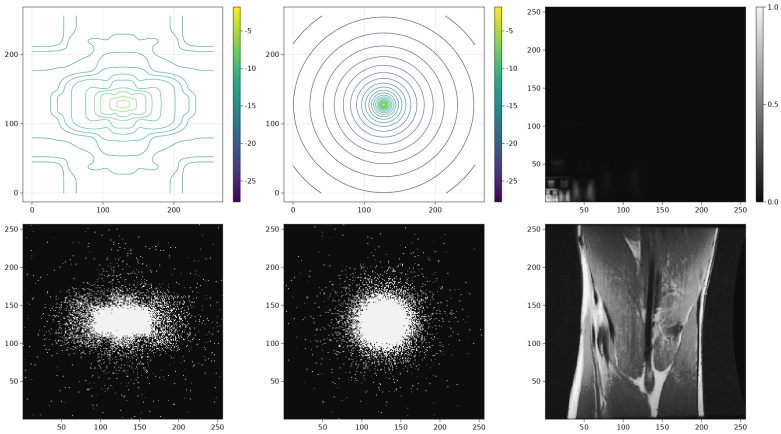

Figure 3.3: Adapted variable density sampling scheme (left column) vs polynomial decay (middle column). Matrix W of sparse support distribution in the DB4 wavelet basis (top right) and test image (bottom right). The resulting PSNR values are: Adapted - 27.9 and Polynomial - 26.8.

the matrix W we again transform each training image into the DB4 wavelet basis and apply a threshold of about 0.006 to get distribution of non-zero coefficients (top right). This time the resulting weights are non-symmetrical and hence plugging them into Formula (3.8) results in a non-symmetrical subsampling density, thereby *adapting* to the underlying structure of the signals. This makes the difference between the adapted subsampling distribution and the polynomial subsampling strategy more pronounced, which will also result in greater differences in the PSNR. Sampling 10% of measurements from the adapted and polynomial densities (bottom left and middle), we get by applying the Nesta algorithm to (3.1) that our adapted subsampling scheme outperforms the heuristically inspired polynomial subsampling strategy — resulting in a PSNR value of 27.9 compared to 26.8.

This difference in performance gets even more pronounced in the next experiment (Figure 3.4), where we use the same setup (and dataset) as in the first experiment, but **flip** the sparse

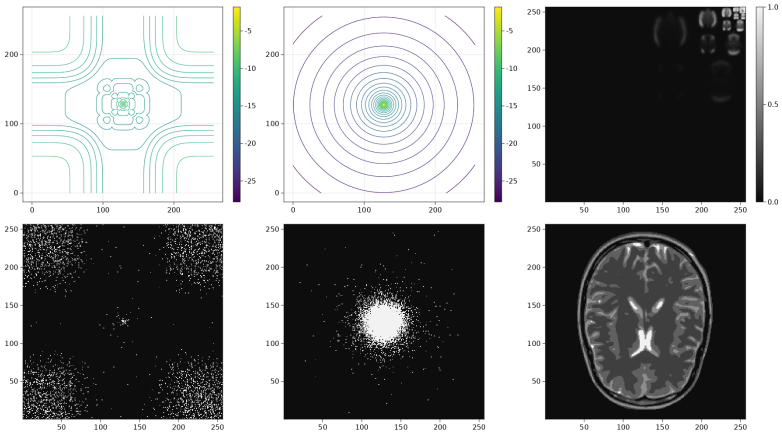

Figure 3.4: Adapted variable density sampling scheme (left column) vs polynomial decay (middle column). Matrix W of sparse support distribution in the DB4 wavelet basis (top right) and test image (bottom right). The resulting PSNR values are: Adapted - 22.9 and Polynomial - 11.6.

coefficients of each image (including the test image) by applying the transform $x \mapsto x^f \in \mathbb{C}^K$, $x_1^f = x_K, x_2^f = x_{K-1}, \cdots, x_K^f = x_1$ to the vectorised sparse coefficients. This is inspired by the so-called flip test [2]. Obviously, the estimated distribution of the sparse supports is now flipped as well and plugging these weights p into Formula (3.8) yields a completely different sampling distribution. We again sample 10% of measurements from the 2D k-space (bottom left and middle). This time, our adapted subsampling strategy easily outperforms the heuristic polynomial decay subsampling strategy — resulting in a PSNR value of 22.7 compared to 12.0.

3.6. Sparsity in levels and blocks of measurements

We now analyse one of the most common framework in modern compressed sensing theory.

3.6.1 Sparsity in levels

A frequent assumption in modern compressed sensing theory is sparsity in levels [2, 14, 3]. To apply our results to this framework we assume that $K = 2^{J+1}$ for some $J \in \mathbb{N}$ and set $A_0 = \mathcal{F}\Psi^*$, where $\mathcal{F}$ is the 1-D Fourier transform with rows indexed from $-K/2+1$ to $K/2$ and Ψ is the 1-D inverse Haar wavelet transform. Denote by Ω the dyadic partition of the set $\{1, \cdots, K\}$ where $\Omega_0 := 0$ and $\Omega_j := \{2^j + 1, \cdots, 2^{j+1}\}$ for $j = 1, \cdots, J$. Further denote by M the $J+1$ frequency bands of the discrete Fourier transform $\mathcal{F}$, i.e., $M_0 := \{0, 1\}$ and $M_j := \{-2^j + 1, \cdots, -2^{j-1}\} \cup \{2^{j-1} + 1, \cdots, 2^j\}$ for $j = 1, \cdots, J$. Lemma 1 in [1] states that for $\ell \in M_i$ and $k \in \Omega_j$

$$|a_{k,\ell}|^2 \lesssim 2^{-j} 2^{|j-i|}. \tag{3.10}$$

We define the **average sparsity in level** ℓ as

$$S_\ell := \|p_{\Omega_\ell}\|_1 \tag{3.11}$$

For simplicity we assume $S_\ell > 1$ for all $1 \leq \ell \leq J$. Plugging this into (3.8) yields for $k \in M_j$

$$a_k D_p a_k^* \lesssim 2^{-j} S_j + 2^{-j} \sum_{p \neq j} 2^{|j-p|} S_p, \tag{3.12}$$

and thus by using ω as defined in (3.8) our main result yields the sufficient condition

$$m \gtrsim \left(\sum_j S_j + \sum_{p \neq j} 2^{|j-p|} S_p \right) \log^3(K/\varepsilon), \tag{3.13}$$

in line with results in [3].

3.6.2 Blocks of measurements

Even though the above sampling strategies yield very good reconstruction results, probing measurements independently at random is infeasible — or at least impractical — in most real applications, see [14] and references therein. Luckily, our results easily extend to the case of blocks of measurements B_k.

Sensing vertical (or horizontal) lines in 2D

We will again follow the notation in [14, 3] very closely to facilitate easier comparison. Assume again that $K = 2^{J+1}$ for some odd $J \in \mathbb{N}$. Let $\Phi \in \mathbb{C}^{\sqrt{K} \times \sqrt{K}}$ be a unitary matrix (for example the discrete 1D Fourier-Haar transform matrix) and assume that our set of possible measurements is given by

$$A_0 = \Phi \otimes \Phi \in \mathbb{C}^{K \times K}, \tag{3.14}$$

where $\otimes$ denotes the Kronecker product. With this notation, we define blocks of measurements which, in a 2D Fourier-Wavelet setting would correspond to vertical lines in frequency space. For this set

$$B_k := \Phi_{k,:} \otimes \Phi = \left(\Phi_{k,1} \Phi \mid \ldots \mid \Phi_{k,\sqrt{K}} \Phi \right) \in \mathbb{C}^{\sqrt{K} \times K} \quad \text{for all} \quad 1 \leq k \leq \sqrt{K}. \tag{3.15}$$

The separable nature of this setup has the big advantage that the matrix $B_k^* B_k$ has a very nice representation. Note that in our main result we have to control $\|B_k D_p B_k^*\| = \|D_{\sqrt{p}} B_k^* B_k D_{\sqrt{p}}\|$. Using that Φ is a unitary matrix we see

$$B_k^* B_k = (\Phi_{k,:} \otimes \Phi)^* (\Phi_{k,:} \otimes \Phi) = (\Phi_{k,:}^* \Phi_{k,:} \otimes \Phi^* \Phi) = (\Phi_{k,:}^* \Phi_{k,:} \otimes \mathbb{I}). \tag{3.16}$$

For our weight vector $p \in \mathbb{R}^K$ we denote by $W \in \mathbb{R}^{\sqrt{K} \times \sqrt{K}}$ the matrix satisfying $\mathrm{vec}(W) = p$. Multiplying $B_k^* B_k = (\Phi_{k,:}^* \Phi_{k,:} \otimes \mathbb{I})$ from the left and right with the diagonal matrix $D_{\sqrt{p}}$ and taking the operator norm yields

$$\|D_{\sqrt{p}} (\Phi_{k,:}^* \Phi_{k,:} \otimes \mathbb{I}) D_{\sqrt{p}}\| = \|D_{\sqrt{p}} \begin{pmatrix} \Phi_{k,1}^* \Phi_{k,1} \mathbb{I} & \ldots & \Phi_{k,1}^* \Phi_{k,\sqrt{K}} \mathbb{I} \\ \vdots & \ddots & \vdots \\ \Phi_{k,\sqrt{K}}^* \Phi_{k,1} \mathbb{I} & \ldots & \Phi_{k,\sqrt{K}}^* \Phi_{k,\sqrt{K}} \mathbb{I} \end{pmatrix} D_{\sqrt{p}}\|. \tag{3.17}$$

Since reordering of columns and rows does not change the operator norm, we apply the permutation $R : \mathbb{K} \mapsto \mathrm{vec}(\mathrm{vec}^{-1}(\mathbb{K})^*)$ to both the columns and rows of the above matrix and set

$p' := R(p)$ to get

$$\|D_{\sqrt{p}}(\Phi_{k,:}^*\Phi_{k,:} \otimes \mathbb{I})D_{\sqrt{p}}\| = \left\| D_{\sqrt{p'}} \begin{pmatrix} \Phi_{k,:}^*\Phi_{k,:} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \Phi_{k,:}^*\Phi_{k,:} \end{pmatrix} D_{\sqrt{p'}} \right\| \tag{3.18}$$

$$= \max_{1\leq \ell \leq \sqrt{K}} \|\Phi_{k,:} D_{W_{\ell,:}}^{1/2}\|_2^2 = \max_{1\leq \ell \leq \sqrt{K}} \sum_{i=1}^{\sqrt{K}} |\Phi_{k,i}|^2 W_{\ell,i}. \tag{3.19}$$

So we look for the row v of the matrix W, such that $\|\Phi_{k,:}D_{\sqrt{v}}\|_2^2$ is maximised. This encapsulates the relationship between the structure of the blocks of measurements and the structure of the sparse signals via their distribution. By the same argument as above we also see that

$$\|B_k^* B_k\|_{\infty,1} = \|\Phi_k\|_\infty^2. \tag{3.20}$$

Plugging this into our formula for blocks (3.6) yields

$$\omega_k := \frac{\max\left\{\max_{1\leq \ell \leq \sqrt{K}} \sum_{i=1}^{\sqrt{K}} |\Phi_{k,i}|^2 W_{\ell,i}, \|\Phi_k\|_\infty^2\right\}}{L}, \tag{3.21}$$

where L is the normalisation constant. If instead of vertical lines one would take horizontal lines

$$B_k := \Phi \otimes \Phi_{k,:}, \tag{3.22}$$

we would get

$$B_k^* B_k = \begin{pmatrix} \Phi_{k,:}^*\Phi_{k,:} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \Phi_{k,:}^*\Phi_{k,:} \end{pmatrix}, \tag{3.23}$$

without any reordering. Hence in this case

$$\|D_{\sqrt{p}} B_k^* B_k D_{\sqrt{p}}\| = \max_{1\leq \ell \leq \sqrt{K}} \sum_{i=1}^{\sqrt{K}} |\Phi_{k,i}|^2 W_{i,\ell}, \tag{3.24}$$

which amounts to taking the maximum over all columns of the matrix W. Plugging this back into our formula for blocks (3.6) yields

$$\omega_k := \frac{\max\left\{\max_{1\leq \ell \leq \sqrt{K}} \sum_{i=1}^{\sqrt{K}} |\Phi_{k,i}|^2 W_{i,\ell}, \|\Phi_k\|_\infty^2\right\}}{L}, \tag{3.25}$$

where L is again the normalisation constant.

Vertical Fourier-Haar lines

We now apply the above analysis to the special case where $\Phi = \mathcal{F}\mathcal{H}^*$ is the 1D Fourier-Haar transform. This yields that A_0 is the separable 2D Fourier-Haar transform[2]. Let $p \in \mathbb{R}^K$ again be our weight vector and define the matrix $W \in \mathbb{R}^{\sqrt{K}\times\sqrt{K}}$ such that $\text{vec}(W) = p$. We again

2. In all other experiments we use non-separable 2D wavelet transforms.

denote by M_ℓ the frequency bands of the one dimensional Fourier transform and by Ω_ℓ the dyadic partition (see previous subsection). In the 2D setting we define the **average sparsity in level** ℓ as

$$S_\ell := \max_k \|W_{k,\Omega_\ell}\|_1. \tag{3.26}$$

This is equivalent to the 1D case up to taking the maximum over all rows of the matrix W. Using (3.10) and assuming that $S_\ell > 1$ for all $1 \leq \ell \leq J$, the above analysis yields for $k \in M_j$

$$\|B_k^* D_p B_k\| = \max_{1\leq \ell \leq \sqrt{K}} \sum_{i=1}^{\sqrt{K}} |\Phi_{k,i}|^2 W_{\ell,i} \leq \sum_{i=1}^{\sqrt{K}} \max_{1\leq \ell \leq \sqrt{K}} |\Phi_{k,i}|^2 W_{\ell,i} \tag{3.27}$$

$$\lesssim 2^{-j} S_j + 2^{-j} \sum_{p\neq j} 2^{|j-p|} S_p, \tag{3.28}$$

and thus by using ω as defined in (3.21) our main result yields the sufficient condition

$$m \gtrsim \left(\sum_j S_j + \sum_{p\neq j} 2^{|j-p|} S_p \right) \log^3(K/\varepsilon), \tag{3.29}$$

in line with results in [3]. Note that the first inequality in (3.27) is rather crude and potentially loses a lot of information about the relationship between the matrix W and the structure of the 2D Fourier-Haar matrix $A_0 = \mathcal{F}\Psi^*$. This is why in our experiments we will stick with the quantity $\|B_k^* D_p B_k\| = \max_{1\leq \ell \leq \sqrt{K}} \sum_{i=1}^{\sqrt{K}} |\Phi_{k,i}|^2 W_{\ell,i}$.

Numerical experiments of blocks of measurements Fourier - DB4

In this subsection we will use blocks of measurements in numerical experiments— Figure 3.5. We conduct two experiments, first by measuring along horizontal lines in the 2D k-space (left column) and then by measuring square blocks of size 16×16 in the 2D k-space (middle column). We again use the Brain dataset with a threshold of around 0.023 to generate a estimate of the matrix W in the *separable* 2D DB4 wavelet basis (top right). Plugging these estimated weights into Formula (3.6) we get an adapted sampling distribution on the vertical lines (top left) and on the square blocks (top middle). Sampling 20% of measurements from the 2D k-space (middle row) we get good reconstruction of the reference image (bottom right) for both measurement techniques (bottom left and middle). This shows how our results also apply to the setting of blocks of measurements.

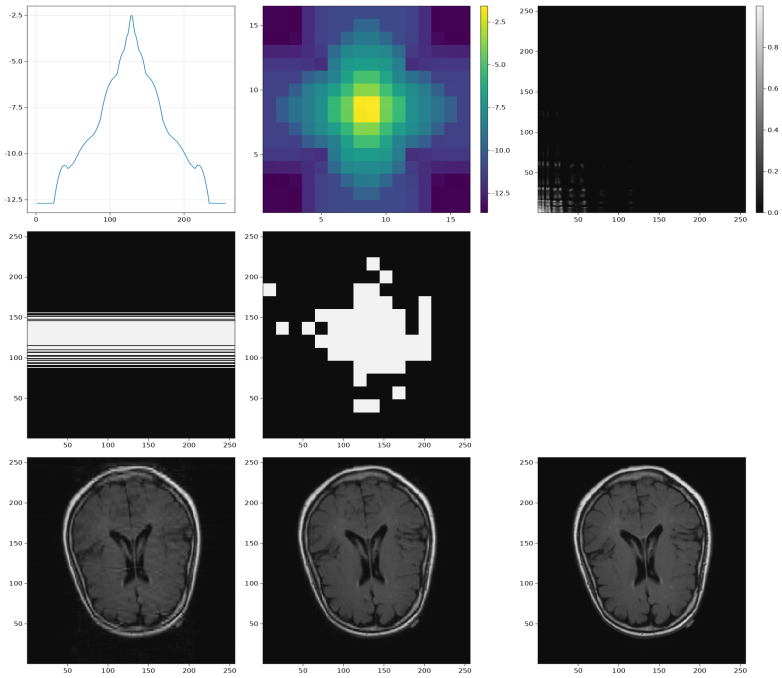

Figure 3.5: Adapted variable density sampling schemes with vertical lines (left column) and squares (middle column). Matrix W of sparse support distribution in the separable 2D DB4 wavelet basis (top right), test image (bottom right) and reconstructions (bottom left and middle). The resulting PSNR values are: Lines - 29.9 and Squares - 33.9.

3.7. Proof of Theorem 3.2

Now we turn to proving Theorem 3.2. Note that we have three sources of randomness: the signs σ, the set of random measurements J and the random supports I. Strictly speaking, we are working on the product measure of the three, but in slight abuse of notation, we will write $\mathbb{P}_\sigma$, $\mathbb{P}_J$ and $\mathbb{P}_S$ to indicate the probability measure that we use for the corresponding concentration inequalities. The exact statement of Theorem 3.2 reads as

Theorem 3.5 *Assume that the signals follow the model in 3.1, where the support $I \subseteq \mathbb{K}$ is chosen according to the rejective sampling model with probabilities $p_1, \cdots, p_K$ such that $\sum_{k=1}^K p_k = S$ and $0 < p_k \leq 1$. If the measurements B_k are sampled according to probabilities ω_k and if*

$$m \geq \max_k \frac{\|B_k^* B_k\|_{\infty,1}}{\omega_k} 128 \log(216 \cdot 6K^2/\varepsilon) \log^2(168K/\varepsilon), \quad \text{and}$$
$$m \geq \max_k \frac{\|B_k D_p B_k^*\|_{2,2}}{\omega_k} 128 e^2 \log^2(168K/\varepsilon), \tag{3.30}$$

then (3.1) *recovers the sparse signal with probability* $1 - \varepsilon$.

Before beginning with the proof, we state 5 concentration inequalities. Recall the definition of the matrices $D_I = \mathbb{I}_I \in \mathbb{R}^{K \times K}$, where $I \subseteq \{1, \cdots, K\}$ with $|I| = S$. Define the quantities

$$\Lambda_I := \max_k \frac{\|D_I B_k^* B_k D_I\|_{2,2}}{\omega_k m} \quad \text{and} \quad \kappa := \max_k \frac{\|B_k^* B_k\|_{\infty,1}}{\omega_k m}.$$

For a fixed support I, the Matrix Bernstein inequality [84] applied to the random matrices $A_I^* A_I - \mathbb{I}$ yields

Lemma 3.6 (Lemma 2.1 [19], Lemma C.1 [14]) *Let I be a fixed support of cardinality S and let A depend on the draw of the j_ℓ. Then for all $t \geq 0$, we have*

$$\mathbb{P}_J \left(\|A_I^* A_I - \mathbb{I}\| \geq t \right) \leq 2S \exp\left(-\frac{t^2/2}{\Lambda_I (1+t)/3} \right).$$

Proof First note that by zero-padding the matrix $A_I^* A_I - \mathbb{I} \in \mathbb{C}^{S \times S}$, we get $\|A_I^* A_I - \mathbb{I}\| = \|D_I A^* A D_I - \mathbb{I}\|$. So to keep the notation uncluttered, we will bound $\|D_I A^* A D_I - \mathbb{I}\|$ as in Chapter 2. Write

$$D_I A^* A D_I - \mathbb{I} = \sum_{k=1}^m \frac{D_I B_{j_k}^* B_{j_k} D_I}{\omega_{j_k} m} - \mathbb{I} = \sum_{k=1}^m \frac{1}{m} \left(\frac{D_I B_{j_k}^* B_{j_k} D_I}{\omega_{j_k}} - \mathbb{I} \right) = \sum_{k=1}^m X_k,$$

where $X_k := \frac{1}{m} \left(\frac{D_I B_{j_k}^* B_{j_k} D_I}{\omega_{j_k}} - \mathbb{I} \right)$. By definition of the j_k, we have $\mathbb{E}[X_k] = 0$. Further

$$\|X_k\|_{2,2} \leq \frac{1}{m} \max\left(\max_k \frac{\|D_I B_k^* B_k D_I\|_{2,2}}{\omega_k} - 1, 1 \right) \leq \Lambda_I.$$

To bound the variance, we note

$$\begin{aligned} 0 \preceq \mathbb{E}[X_k^2] &= \mathbb{E}\left[\left(\frac{D_I B_{j_k}^* B_{j_k} D_I}{\omega_{j_k} m} \right)^2 \right] - \frac{1}{m^2} \mathbb{I} \\ &\preceq \Lambda_I \mathbb{E}\left[\frac{D_I B_{j_k}^* B_{j_k} D_I}{\omega_{j_k} m} \right] \preceq \Lambda_I \frac{1}{m} \mathbb{I}, \end{aligned}$$

which leads to $\sigma^2 = \| \sum_{k=1}^m \mathbb{E}[X_k^2] \|_{2,2} \leq \Lambda_I$. An application of the Matrix Bernstein inequality yields the result. ∎

Further, for I fixed, and $i \in I^c$, we are going to apply the vector Bernstein inequality [52] to $\|A_I^* A_i\|_2$. Together with a union bound this yields

Lemma 3.7 *Let I be a fixed support of cardinality S and let A depend on the draw of the j_ℓ. Then for all $t > 0$, we have*

$$\mathbb{P}_J \left(\max_{i \in I^c} \|A_I^* A_i\|_2 \geq t \right) \leq 28K \exp\left(-\frac{t^2/2}{\Lambda_I + \sqrt{\Lambda_I \kappa} \cdot t/3} \right).$$

Proof Fix $i \in I^c$. Again by zero-padding, we have

$$\|A_I^* A e_i\|_2 = \| \sum_{k=1}^m \frac{1}{m} \frac{D_I B_{j_k}^* B_{j_k} e_i}{\omega_{j_k}} \|_2 = \| \sum_{k=1}^m X_k \|_2,$$

where $X_k := \frac{1}{m} \frac{D_I B_{j_k}^* B_{j_k} e_i}{\omega_{j_k}}$. Since $i \in I^c$, we have $\mathbb{E}[X_k] = \frac{1}{m} D_I \sum_{\ell=1}^M B_\ell B_\ell^* e_i = \frac{1}{m} P_I e_i = 0$. Further

$$\max_k \|X_k\|_2 = \max_k \left\| \frac{D_I B_k^* B_k e_i}{\omega_k m} \right\|_2 \leq \sqrt{\Lambda_I \kappa}$$

To bound the variance, note that

$$\mathbb{E}[\|X_k\|_2^2] = \mathbb{E}\left[\left\|\frac{D_I B_{j_k}^* B_{j_k} e_i}{\omega_{j_k} m}\right\|_2^2\right] \leq$$
$$\leq \Lambda_I \mathbb{E}\left[\left\|\frac{B_{j_k} e_i}{\sqrt{\omega_{j_k} m}}\right\|_2^2\right] = \Lambda_I \|e_i\|_2^2 \frac{1}{m} = \frac{\Lambda_I}{m}.$$

This leads to $\sigma^2 = \sum_{k=1}^m \mathbb{E}[\|X_k\|_2^2] \leq \Lambda_I$. A union bound finishes the proof. ■

We further use the following Hoeffding-like tail bound for sums of centered complex random variables — see ([35] Corollary 7.21 and Corollary 8.10).

Lemma 3.8 *Let $M \in \mathbb{C}^{K\times S}$ be a matrix and $x \in \mathbb{R}^S$ such that $\operatorname{sign}(x_i) \in \mathbb{R}^S$ is an independent Rademacher sequence. Then, for all $t \geq 0$*

$$\mathbb{P}_\sigma\left(\|Mx\|_\infty \geq t\right) \leq 2K \exp\left(-\frac{t^2}{2\|M\|_{\infty,2}^2 \|x\|_\infty^2}\right).$$

The key ingredient to prove Theorem 3.2 is the following concentration inequality for the operator norm of random submatrices with non-uniformly distributed supports which can be found in [67] and Chapter 2. This is what allows us to go one step further than existing results in analysing the underlying relationship between the sensing matrix and the distribution of sparse supports. For convenience, we restate the result.

Lemma 3.9 ([67]) *Let $H \in \mathbb{C}^{K\times K}$ be a matrix with zero diagonal. Assume that the support $I \subseteq \mathbb{K}$ is chosen according to the rejective sampling model with probabilities $p_1, \ldots, p_K$ such that $\sum_{i=1}^K p_i = S$. Further let p denote the corresponding weight vector. If $t \geq 2e^2 \|D_{\sqrt{p}} H D_{\sqrt{p}}\|$ and*

$$\|H\|_{\infty,1} \leq \frac{t}{2\log(216K/\varepsilon)}$$
$$\|HD_{\sqrt{p}}\|_{\infty,2}^2 \leq \frac{t^2}{4e^2\log(216K/\varepsilon)},$$

then $\mathbb{P}_S(\|D_I H D_I\| \geq t) \leq \varepsilon$.

Now we are finally able to state the proof of Theorem 3.2.

Proof From [81, 36] we know that if $\|A_{I^c}^* A_I (A_I^* A_I)^{-1} \sigma_I\|_\infty < 1$, then x is the unique solution of the ℓ_1-minimisation problem (3.1). Set $M := A_{I^c}^* A_I (A_I^* A_I)^{-1}$ and assume that $\vartheta_I := \|A_I^* A_I - \mathbb{I}\| \leq 1/2$. Then

$$\|M\|_{\infty,2} = \|A_{I^c}^* A_I (A_I^* A_I)^{-1}\|_{\infty,2} \leq \|A_{I^c}^* A_I\|_{\infty,2} \|(A_I^* A_I)^{-1}\|_{2,2} \leq 2\|A_{I^c}^* A_I\|_{\infty,2}.$$

Noting that $\|A_{I^c}^* A_I\|_{\infty,2} = \max_{i\in I^c} \|A_I^* A_i\|_2$ we have

$$\mathbb{P}\left(\|M\sigma\|_\infty \geq 1\right) \leq \mathbb{P}_\sigma\left(\|M\sigma\|_\infty \geq 1 \mid \|M\|_{\infty,2} \leq 2\gamma\right)$$
$$+ \mathbb{P}\left(\|A_I^* A_I - \mathbb{I}\| \geq 1/2\right) + \mathbb{P}\left(\max_{i\in I^c} \|A_I^* A_i\|_2 \geq \gamma\right)$$

Setting $\gamma^2 = \frac{1}{8\log(6K/\varepsilon)}$ and applying Lemma 3.8 to $M\sigma$ yields that the first term on the right hand side is bound by $\varepsilon/3$. Further

$$\begin{aligned}
&\mathbb{P}\left(\|A_I^* A_I - \mathbb{I}\| \geq 1/2\right) + \mathbb{P}\left(\max_{i\in I^c} \|A_I^* A_i\|_2 \geq \gamma\right) \\
&\quad \leq \mathbb{P}_J\Big(\|A_I^* A_I - \mathbb{I}\| \geq 1/2 \Big| \Lambda_I \leq v\Big) + \mathbb{P}_S(\Lambda_I \geq v) \\
&\qquad + \mathbb{P}_J\left(\max_{i\in I^c} \|A_I^* A_i\|_2 \geq \gamma \middle| \Lambda_I \leq v\right) + \mathbb{P}_S\left(\Lambda_I \geq v\right)
\end{aligned}$$

Setting $v := \frac{1}{32\log^2(168K/\varepsilon)}$ and using that by the assumptions in Theorem 3.2

$$\kappa \leq \frac{1}{128\log(256\cdot 6K^2/\varepsilon)\log^2(168K/\varepsilon)},$$

an application of Lemma 3.6 and Lemma 3.7 together with the observation that $v + \sqrt{v\kappa}\cdot\gamma/3 \leq \gamma^2$ yields

$$\begin{aligned}
&\mathbb{P}_J\Big(\|A_I^* A_I - \mathbb{I}\| \geq 1/2 \Big| \Lambda_I \leq v\Big) + \mathbb{P}_J\left(\max_{i\in I^c} \|A_I^* A_i\|_2 \geq \gamma \middle| \Lambda_I \leq v\right) \\
&\quad \leq 2S\exp\left(-\frac{1/8}{v(1+1/2)/3}\right) + 28K\exp\left(-\frac{\gamma^2/2}{v+\sqrt{v\kappa}\cdot\gamma/3}\right) \\
&\quad \leq 2S\exp\left(-\frac{1}{4v}\right) + 28K\exp\left(-\frac{\gamma^2}{4v}\right) \\
&\quad \leq \varepsilon/3.
\end{aligned}$$

So to finish the proof we have to show that $\mathbb{P}(\Lambda_I \geq v) \leq \varepsilon/6$. To that end define the matrices

$$H_k := \frac{B_k^* B_k - \operatorname{diag}(B_k^* B_k)}{\omega_k m}.$$

Recall that by definition,

$$\Lambda_I := \max_k \frac{\|D_I B_k^* B_k D_I\|_{2,2}}{\omega_k m} \quad \text{and} \quad \kappa := \max_k \frac{\|B_k^* B_k\|_{\infty,1}}{\omega_k m}.$$

By our assumptions, we have

$$\Lambda_I \leq \max_k \|D_I H_k D_I\| + \|\operatorname{diag}\left(\frac{D_I B_k^* B_k D_I}{\omega_k m}\right)\| \leq \max_k \|D_I H_k D_I\| + \kappa \leq \max_k \|H_k\| + v/2.$$

So we have to show that $\mathbb{P}(\max_k \|D_I H_k D_I\| \geq v/2) \leq \varepsilon/6$, which we will do by showing that $\mathbb{P}(\|D_I H_k D_I\| \geq v/2) \leq \varepsilon/(6K)$ together with a union bound. By applying 3.9 to each H_k this is satisfied, if

$$\begin{aligned}
\|H_k D_{\sqrt{p}}\|_{\infty,2}^2 &\leq \frac{(v/2)^2}{4e^2\log(216\cdot 6K^2/\varepsilon)} \\
\|H_k\|_{\infty,1} &\leq \frac{v/2}{2\log(216\cdot 6K^2/\varepsilon)},
\end{aligned}$$

and $v \geq 2e^2\|D_{\sqrt{p}} H_k D_{\sqrt{p}}\|$. Using that

$$\|H_k D_{\sqrt{p}}\|_{\infty,2}^2 \leq \frac{\|B_k^* B_k D_{\sqrt{p}}\|_{\infty,2}^2}{\omega_k m} \leq \frac{\|B_k^*\|_{\infty,2}^2 \|B_k D_{\sqrt{p}}\|_{2,2}^2}{\omega_k^2 m^2} \leq \kappa \max_k \frac{\|B_k D_p B_k^*\|}{\omega_k m},$$

and $\|H_k\|_{\infty,1} \leq \kappa$, this follows from the assumptions in Theorem 3.2.

Remark 3.10 *The proof of our main result relies heavily on the random signs of our signals. One could remove this assumption by instead employing the so-called "golfing scheme" proposed in [40]. Following the argument in [19] one should be able to derive similar results in the case of deterministic sign patterns. Since this would not have any impact on the optimal sampling distribution we opted for the shorter proof presented here.*

■

3.8. Discussion

The above results showed that the optimal variable density subsampling strategy in a compressed sensing setup should not only depend on the structure of the sensing and sparsity matrices, but also on the distribution of sparsity patterns of the signals to be measured. We derived lower bounds on the number of measurements to ensure recovery of the sparse signals with high probability and derived a simple formula for the optimal subsampling strategy. We showed that this distribution can be estimated from a training set and that the resulting adapted subsampling scheme provides state of the art performance in a range of situations. For future work it would be interesting to analyse different settings of blocks of measurements, where explicit lower bounds on the number of measurements can be derived. One of the main assumptions in the chapter is the existence of a basis such that the signals can be sparsely represented. How to learn such a basis from data for a general signal class will be the topic of the next chapter.

Chapter 4

Dictionary learning convergence

In the previous chapter we have seen that the assumption of a sparsifying basis is crucial for compressed sensing. In some applications such a basis is unknown and thus has to be learned from data via dictionary learning. In this chapter we derive sufficient conditions for convergence of two of the most popular dictionary learning algorithms - Method of Optimal Directions (MOD) and Approximate K-SVD (aK-SVD). We show that given a well-behaved initialisation that is either within distance at most $1/\log(K)$ to the generating dictionary or has a special structure ensuring that each element of the initialisation only points to one generating element, then they will converge with geometric convergence rate to the generating dictionary.

4.1. Introduction

Dictionary learning tries to find structure in data by decomposing a data matrix $Y = (y_1, \cdots, y_N)$, where $y_i \in \mathbb{R}^d$, into the product of a dictionary matrix $\Phi = (\phi_1, \cdots, \phi_K) \in \mathbb{R}^{d\times K}$ and a sparse coefficient matrix $X = (x_1, \cdots, x_N) \in \mathbb{R}^{K\times N}$ such that

$$Y \approx \Phi X \quad \text{and} \quad X \text{ sparse.} \tag{4.1}$$

There exist many algorithms to choose from when trying to tackle the above problem [34, 6, 33, 72, 47, 49, 50, 76, 51, 68, 54] and a growing number of theoretical results to accompany them [38, 77, 4, 7, 69, 70, 39, 9, 8, 72, 78, 78, 5, 20, 61]. We point the interested reader to the surveys [65, 71] for easy access into the world of dictionary learning. The common starting point of many dictionary learning algorithms is to formulate the problem in 4.1 as a minimisation problem,

$$\text{argmin}_{\Psi,X} \|Y - \Psi X\|_F^2 \qquad \text{s.t.} \quad \|\psi_k\|_2 = 1 \quad \text{and} \quad \|x_n\|_0 \leq S., \tag{4.2}$$

where S is a prescribed sparsity level and the condition $\|\psi_k\|_2 = 1$ prevents scaling ambiguities between the dictionary and the sparse codes x_x. Since the problem is highly non-convex with many equivalent global minima corresponding to different orderings and signs of the dictionary and even the full gradient cannot be calculated explicitly, one usually one tries to find a solution by alternate minimisation — meaning iteratively fixing the sparse codes X and updating the dictionary Ψ and vice versa. Popular alternate minimisation algorithms include MOD (Method of Optimal Directions) [33], K-SVD (K Singular Value Decompositions) [6] and ITKrM (Iterative Thresholding and K residual Means) [72]. Despite their popularity,

the main drawback of these algorithms is the relative lack of theoretical results underpinning their empirical success. To the best of our knowledge, there exist no recovery guarantees for the K-SVD algorithm. For MOD the situation is a little different as local convergence for an alternate minimisation algorithm identical to MOD was proven in [5]. The only difference between their proposed algorithm and the original MOD is the sparse approximation step, which uses ℓ_1-minimisation instead of OMP. The main drawback of this result is that it works only for initialisations with maximal atom-wise ℓ_2-distance $1/S^2$ from the ground truth, which limits its practical relevance considerably. For ITKrM a recent result showed **contraction** under very relaxed conditions [56]. While the result does not guarantee convergence it shows that the algorithm contracts towards the generating dictionary, if the current guess it is either in a ball of radius $1/\sqrt{\log(K)}$ around the solution or has a special type of structure where each atom corresponds only to one atom of the ground truth meaning that no two estimated atom point to the same generating atom and there is sufficient separation. In this case the distance between ground truth and initialisation might be close to the theoretical limit $\sqrt{2}$, improving considerably upon previous theoretical results.
The holy grail in dictionary learning theory are global convergence guarantees, but the highly non-convex nature of the dictionary learning problem makes such results prohibitively hard and most likely impossible for practically usable algorithms based on alternating minimisation, which in simulations can get trapped in spurious saddle points, 5. There do exist algorithms with global recovery guarantees [29, 4, 9], but either their prohibitively high computational complexity or numerical sensitivity make them rather hard to recommend in a practise.

4.1.1 Our Contribution

In this chapter we will analyse a slight adaptation of the approximate K-SVD (aK-SVD) [64], which exchanges the costly SVD in the dictionary update step of the original K-SVD algorithm for a power iteration and OMP in the sparse approximation step for the computationally lighter Thresholding algorithm. We will also analyse the MOD where we again use Thresholding as the sparse approximation step. We show that both algorithms **converge** to the generating dictionary under very general assumptions similar in spirit to the conditions in [56] meaning that for a well-behaved initialisation it either has to be in a ball of radius $1/\sqrt{\log(K)}$ or have such a structure that each element of the initialisation only points to one generating element, ensuring sufficient separation for the sparse approximation step not to mix things up. If these criteria on the structure are met, the distance may be close to the theoretical limit of $\sqrt{2}$ while convergence is still guaranteed by our result. Further we will make use of the same non-uniform signal model used throughout this thesis, allowing us to use model situations, where some atoms in the generating dictionary are used more frequently than others. This in itself represents a great generalisation of existing results, where the sparse supports are usually assumed to be chosen uniformly at random among all possible subsets of cardinality S.

4.1.2 Organisation

In the first section, we define the signal model and special notation that will be used in this chapter. In Section 4.3 we will recall the dictionary learning algorithms that will be analysed in this chapter. We state our main result in Section 4.4 and prove it in Section 4.5. To make the proof more accessible, we defer the technical results it relies on to Section 4.6, Section 4.7, Section 4.8 and, in particular, Section 4.9, which contains the results necessitated by the non-uniform support model.

4.2. Setting

We define the following model for our signals.

Definition 4.1 (Signal model) *Given a generating dictionary $\Phi \in \mathbb{R}^{d\times K}$ consisting of K normalized atoms, we model our signals as*

$$y = \Phi_I x_I = \sum_{i\in I} \phi_i x_i, \quad x_i = c_i \sigma_i, \tag{4.3}$$

where the support $I = \{i_1, \dots i_S\} \subseteq \mathbb{K}$ is chosen according to the rejective sampling model with probabilities $p_1, \dots, p_K$ such that $\sum_{i=1}^K p_i = S$ and $0 \leq p_k \leq 1/6$, the coefficient sequence $c = (c_i)_i \in \mathbb{R}^K$ consists of i.i.d. bounded random variables c_i with $0 \leq c_{\min} \leq c_i \leq c_{\max} \leq 1$ and the sign sequence $\sigma \in \{-1, 1\}^K$ is a Rademacher sequence, i.e. its entries σ_i are i.i.d with $\mathbb{P}(\sigma_i = \pm 1) = 1/2$. Supports, coefficients and signs are modeled as independent and we can write $x = \mathbf{1}_I \odot c \odot \sigma$.

The assumption $p_i \leq 1/6$ is to ensure that p_i and the corresponding inclusion probabilities of the rejective sampling model π_i are not too different (see Theorem 4.17). We define the vectors $\alpha, \beta \in \mathbb{R}^K$ via

$$\alpha_i := \langle \phi_i, \psi_i \rangle \qquad \beta_i := \mathbb{E}[c_i^2]$$

and the corresponding diagonal matrices D_α, D_β. We define the distance between the generating dictionary Φ and a guess Ψ as

$$\delta(\Psi, \Phi) := \max\Big\{ \|(\Psi - \Phi) D_{\sqrt{\pi}}\|_{2,2}, \|\Psi - \Phi\|_{2,1} \Big\}.$$

This might not seem intuitive at first glance, but if we are able to show convergence of this distance we are able to also control the weighted operator norm of the error in each step. Sometimes we will also need just the ℓ_2-distance between two dictionary elements

$$\varepsilon(\Psi, \Phi) := \|\Psi - \Phi\|_{2,1} = \max_i \|\psi_i - \phi_i\|_2.$$

If it is clear from context, we will sometimes write ε and δ instead of $\varepsilon(\Psi, \Phi)$ and $\delta(\Psi, \Phi)$. A very important variable which will be used very frequently throughout this chapter is

$$Z := \Phi - \Psi,$$

which is the difference matrix between the generating dictionary Φ and our current dictionary Ψ.

4.3. Algorithms

We will analyse the convergence of two algorithms, K-SVD and MOD, or more accurately slightly modified versions that use Thresholding in the sparse approximation step rather that OMP.

We start with a detailed description of the K-SVD version we will analyse. In the original K-SVD algorithm the dictionary is updated atom by atom with the goal of reducing the part of current error $\|Y - \Psi\hat{X}\|$, which is related to the atom at hand. Concretely to update the k-th atom, based on the current guess for dictionary and coefficients $(\Psi, \hat{X})$, it first sets $\psi_k = 0$,

$$J := \{n : \hat{x}_n(k) \neq 0\} \quad \text{and} \quad E := Y_J - \Psi \hat{X}_J, \tag{4.4}$$

solves via singular value decomposition

$$\text{argmin}_{v,g} \|E - vg^*\|_F. \tag{4.5}$$

and updates $\hat{\psi}_k = v/\|v\|$ and $\hat{X}_{k,J}^{\text{new}} = \|v\| \cdot g^*$. One drawback of this update is that finding the largest left-singular vector of E for each atom in each update step is computationally quite expensive. This is why an approximate version, called aK-SVD, of the popular algorithm was proposed in [64]. There the singular value decomposition in each step is replaced with a simple power iteration using as initial guess for $g = (\hat{X}_{k,J})^*$. This means that

$$\hat{\psi}_k := E(\hat{X}_{k,J})^*/\|E(\hat{X}_{k,J})^*\|_2, \tag{4.6}$$

while the sparse codes in $\hat{X}_{k,J}$ are updated by $(\hat{\psi}_k)^* E$. Looking more closely at the atom update, we can write the updated atom before normalisation as

$$\begin{aligned}\tilde{\psi}_k = E(\hat{X}_{k,J})^* &= (Y - \Psi\hat{X})_J(\hat{X}^*)_{J,k} \\ &= \sum_{n\in J}(y_n - \Psi\hat{x}_n)\hat{x}_n(k) = \sum_{n=1}^{N}[y_n\hat{x}_n(k) - \Psi\hat{x}_n\hat{x}_n(k)],\end{aligned}$$

where in the last inequality we used that $\hat{x}_n(k) = 0$ for $k \notin J$. Recall that first step in the atom update was to set $\psi_k = 0$. Alternatively, if we do not set $\psi_k = 0$, we can write

$$\tilde{\psi}_k = \sum_{n=1}^{N}\left[y_n\hat{x}_n(k) - \Psi\hat{x}_n\hat{x}_n(k) + \psi_k\hat{x}_n(k)^2\right]. \tag{4.7}$$

Now whether the atoms are updated according to (4.5) or (4.3.1) an important part of all K-SVD variants is that the atoms are updated consecutively and together with the corresponding sparse codes. This makes them prohibitively difficult to analyse, as it is not clear how one should order the atoms and each ordering will result in a slightly different learned dictionary. In order to get a more analysable algorithm with less dependencies we therefore propose a slightly simplified version of this algorithm, which we again call approximate K-SVD (aK-SVD), where the sparse codes stay unaltered during the dictionary update step. This comes with the added benefit of the resulting algorithm being parallelisable since each atom update is independent of the others. Further, it allows us to write the unnormalised dictionary update step in closed form as

$$\begin{aligned}\text{aK-SVD:} \qquad \tilde{\Psi} &= \sum_{n=1}^{N}[y_n\hat{x}_n^* - \Psi\hat{x}_n\hat{x}_n^* + \Psi\,\text{diag}(\hat{x}_n\hat{x}_n^*)] \\ &= Y\hat{X}^* - \Psi\hat{X}\hat{X}^* + \Psi\,\text{diag}(\hat{X}\hat{X}^*).\end{aligned} \tag{4.8}$$

A detailed description of the aK-SVD algorithm with Thresholding, that we are going to analyse, can be found in Algorithm 4.3.1.
As can be seen there apart from using Thresholding rather than OMP in the sparse approximation step and updating all atoms at once, we have included another small modification. To avoid instability due to thresholding recovering an incorrect and very ill conditioned support we use a cut-off that sets coefficients that are very large compared to the signal size to zero.
We can also see that aK-SVD is remarkably similar to the ITKrM algorithm, [72]. We only need to replace $[a_n + \psi_k\hat{x}_n(k)] \cdot \hat{x}_n(k)$ by $[a_n + \psi_k(y_n^*\psi_k)] \cdot \text{sign}(y_n^*\psi_k)$ to arrive at the update

Algorithm 4.3.1: Approximate K-SVD (one iteration)

Input : Ψ, Y, S, κ
foreach n **do**
  Initialise $\tilde{\Psi} = 0$, $\hat{X} = 0$;
  $\hat{I}_n = \operatorname{argmax}_{I:|I|=S} \|\Psi_I^* y_n\|_1$; // thresholding
  $\hat{x}_n(\hat{I}_n) = \Psi_{\hat{I}_n}^{\dagger} y_n$; // coefficient estimation
  if $\|\hat{x}_n\|_2 \geq \kappa \|y_n\|_2$ **then**
    $\hat{x}_n = 0$; // set pathological estimates to zero
  end
  $a_n = y_n - \Psi \hat{x}_n$;
  foreach $k \in \hat{I}_n$ **do**
    $\tilde{\psi}_k \leftarrow \tilde{\psi}_k + [a_n + \psi_k \hat{x}_n(k)] \cdot \hat{x}_n(k)$; // dictionary update
  end
end
$\hat{\Psi} \leftarrow (\tilde{\psi}_1/\|\tilde{\psi}_1\|_2, \cdots, \tilde{\psi}_K/\|\tilde{\psi}_K\|_2)$; // atom normalisation
Output: $\hat{\Psi}$

step of the ITKrM algorithm.

The second dictionary learning algorithm we will analyse in this chapter is a variant of the MOD algorithm 4.3.2. As in the aK-SVD algorithm, we will employ the Thresholding algorithm rather than FOCUS in the sparse approximation step. The justification for this can be found in Chapter 5. Now the dictionary update step is where it differs from the above algorithm. Even though both of them try to minimise the ℓ_2-cost of the minimisation problem (4.2), MOD updates all atoms at once, by solving a simplified minimisation problem, where Ψ need not have normalised columns,

$$\tilde{\Psi} = \operatorname{argmin}_{\Psi} \|Y - \Psi \hat{X}\|_F^2, \tag{4.9}$$

and normalising afterwards. The advantage is that the problem above has a closed form solution $\tilde{\Psi} = Y\hat{X}^{\dagger}$, which in case $\hat{X}\hat{X}^*$ has full rank simplifies further to

$$\text{MOD:} \qquad \tilde{\Psi} = Y\hat{X}^*(\hat{X}\hat{X}^*)^{-1} = \sum_{n=1}^{N} y_n \hat{x}_n^* \left(\sum_{n=1}^{N} \hat{x}_n \hat{x}_n^* \right)^{-1}. \tag{4.10}$$

A detailed description of the MOD algorithm with Thresholding can be found in Algorithm 4.3.2. As we can see the dictionary update of both aK-SVD and MOD involves the matrices $\hat{X}\hat{X}^*$ and $Y\hat{X}^*$ or taking into account that $Y = \Phi X$, rather the matrices $\hat{X}\hat{X}^*$ and $X\hat{X}^*$,

$$\begin{aligned} \text{aK-SVD:} \qquad & \tilde{\Psi} = Y\hat{X}^* - \Psi\hat{X}\hat{X}^* + \Psi \operatorname{diag}(\hat{X}\hat{X}^*) = \Phi X\hat{X}^* - \Psi\hat{X}\hat{X}^* + \Psi \operatorname{diag}(\hat{X}\hat{X}^*), \\ \text{MOD:} \qquad & \tilde{\Psi} = Y\hat{X}^*(\hat{X}\hat{X}^*)^{-1} = \Phi X\hat{X}^*(\hat{X}\hat{X}^*)^{-1}. \end{aligned}$$

Now the key to proving convergence is to show that both these matrices or scaled versions of them are essentially diagonal matrices, that is, we define

$$A := \frac{1}{N} X\hat{X}^* = \frac{1}{N} \sum_{n=1}^{N} x_n \hat{x}_n^* \quad \text{and} \quad B := \frac{1}{N} \hat{X}\hat{X}^* = \frac{1}{N} \sum_{n=1}^{N} \hat{x}_n \hat{x}_n^*. \tag{4.11}$$

Algorithm 4.3.2: Method of Optimal Directions (one iteration)

Input : Ψ, Y, S, κ

foreach n **do**

  $\hat{I}_n = \operatorname{argmax}_{I:|I|=S} \|\Psi_I^* y_n\|_1$; `// thresholding`

  $\hat{x}_n(\hat{I}_n) = \Psi_{\hat{I}_n}^{\dagger} y_n$; `// coefficient estimation`

  if $\|\hat{x}_n\|_2 \geq \kappa \|y_n\|_2$ **then**

    $\hat{x}_n = 0$; `// kill pathological estimates`

  end

end

$\hat{X} \leftarrow (\hat{x}_1, \cdots, \hat{x}_N)$;

$\tilde{\Psi} \leftarrow (Y\hat{X}^{\dagger})$; `// dictionary update`

$\hat{\Psi} \leftarrow (\tilde{\psi}_1/\|\tilde{\psi}_1\|_2, \cdots, \tilde{\psi}_K/\|\tilde{\psi}_K\|_2)$; `// atom normalisation`

Output: $\hat{\Psi}$

We first take a closer look at the terms within the sums above, where for simplicity we drop the index n. Assuming that Thresholding finds the correct support, we can write $x\hat{x}^*$ using the zero-padding operator R_I^* as,

$$x\hat{x}^* = x(R_I^*\Psi_I^{\dagger}y)^* = xx^*\Phi^*(\Psi_I^{\dagger})^* R_I,$$

Further assuming that Ψ_I is well conditioned meaning, $\Psi_I^*\Psi_I \approx \mathbb{I}$, we can next approximate $\Psi_I^{\dagger} \approx \Psi_I^*$ leading to

$$x\hat{x}^* \approx xx^*\Phi^*\Psi_I R_I = xx^*\Phi^*\Psi R_I^* R_I = xx^*\Phi^*\Psi \operatorname{diag}(\mathbf{1}_I).$$

As we modelled the generating coefficients as $x = c \odot \sigma \odot \mathbf{1}_1$, using the independence of c, σ, I we get that in expectation over c, σ,

$$\mathbb{E}_{c,\sigma}[xx^*] = \mathbb{E}_c[cc^*] \odot \mathbb{E}_\sigma[\sigma\sigma^*] \odot (\mathbf{1}_I\mathbf{1}_I^*) = D_\beta \operatorname{diag}(\mathbf{1}_I).$$

So A the empirical estimator for $\mathbb{E}[x\hat{x}]$ should we well approximated by

$$A = \frac{1}{N}\sum_{n=1}^{N} x_n\hat{x}_n^* \approx \mathbb{E}[x\hat{x}^*] \approx \mathbb{E}D_\beta \operatorname{diag}(\mathbf{1}_I)\Phi^*\Psi \operatorname{diag}(\mathbf{1}_I) = (D_\beta\Phi^*\Psi) \odot \mathbb{E}[\mathbf{1}_I\mathbf{1}_I^*].$$

The matrix $\mathbb{E}[\mathbf{1}_I\mathbf{1}_I^*]$ simply stores as ij-th entry how often $\{i, j\} \subseteq I$, meaning the diagonal entries ar far larger than the off-diagonal ones, and we can approximate $\mathbb{E}[\mathbf{1}_I\mathbf{1}_I^*] \approx D_\pi + \pi\pi^* \approx D_\pi$. Finally using that $D_\alpha = \operatorname{diag}(\Phi^*\Psi)$ we see that the matrix A should be very close to

$$A \approx \mathbb{E}[x\hat{x}^*] \approx (D_\beta\Phi^*\Psi) \odot D_\pi = D_{\pi\cdot\alpha\cdot\beta} \tag{4.12}$$

Using similar arguments as above we get that B the empirical estimator for $\mathbb{E}[\hat{x}\hat{x}^*]$ should be close to

$$B = \frac{1}{N}\sum_{n=1}^{N} \hat{x}_n\hat{x}_n^* \approx \mathbb{E}[\hat{x}\hat{x}^*] \approx D_{\pi\cdot\alpha^2\cdot\beta}. \tag{4.13}$$

So before normalisation the dictionary updated via aK-SVD should be approximately

$$\tilde{\Psi} = \Phi X\hat{X}^* - \Psi\hat{X}\hat{X}^* + \Psi \operatorname{diag}(\hat{X}\hat{X}^*) = \Phi A - \Psi[B - \operatorname{diag}(B)] \approx \Phi D_{\pi\cdot\alpha\cdot\beta}$$

and the dictionary updated via MOD

$$\tilde{\Psi} = \Phi X \hat{X}^* (\hat{X}\hat{X}^*)^{-1} = \Phi A B^{-1} \approx \Phi D_\alpha^{-1}. \tag{4.14}$$

This means that the output of both dictionary update steps after normalisation should be a dictionary, which is very close to the generating dictionary, and the work to be done in the proof essentially boils down to quantifying the error in the approximation steps outlined above. Concretely, we will prove the following main result.

4.4. Main results

We are finally in a position to state our main result. After that, we will analyse each condition and explain the intuition behind some of the assumptions.

Theorem 4.2 *Assume our signals follow the signal model in 4.1 with probabilities $p_1, \cdots, p_K$ such that $\sum_i p_i = S$ and $0 \leq p_i \leq \frac{1}{6}$. Let $\pi_i := \mathbb{P}_S(i \in I)$ and denote by Φ the generating dictionary and by Ψ the initial guess. Set*

$$\underline{\alpha} := \min_k |\langle \psi_k, \phi_k \rangle| = 1 - \frac{\varepsilon^2}{2} \quad \text{and} \quad \gamma := \frac{c_{\min}}{c_{\max}} \quad \text{and} \quad \rho := 2\kappa^2 \|\Phi\|^2 S \gamma^{-2} \underline{\alpha}^{-2} \underline{\pi}^{-3/2}.$$

We denote by $\delta_\star$ the desired recovery accuracy and assume $\delta_\star \log(K\rho/\delta_\star) \leq \gamma^2/(8C)$. If

$$\max\left\{ \|\Phi D_{\sqrt{\pi}}\|^2, \|\Psi D_{\sqrt{\pi}}\|^2, \mu(\Phi), \mu(\Psi, \Phi)^2, \mu(\Psi) \right\} \leq \frac{\underline{\alpha}^2 \gamma^2}{C^2 \log(K\rho/\delta_\star)} \tag{4.15}$$

for a universal constant C, and if we are given N fresh signals each iteration, then with high probability both MOD and aK-SVD converge with geometric rate up to precision $\delta_\star$ to the generating dictionary Φ. The failure probability in each step is bounded by

$$60K \exp\left(-\frac{N(\delta_\star/32)^2}{2\rho^2 + \rho\delta_\star/32} \right). \tag{4.16}$$

This theorem might seem overwhelmingly complicated at first glance, which is why we will provide some explanations as to what the different conditions and bounds actually mean in practice.

Initialisation: The above result depends on the distance ε between the generating dictionary Φ and the initialisation Ψ via the quantity $\underline{\alpha} = 1 - \varepsilon^2/2$ and the structure of the initialisation via the cross-coherence $\mu(\Psi, \Phi) = \max_{i \neq j} |\langle \psi_i, \phi_j \rangle|$. Assuming for a moment that we are in a uniform support model, i.e. $p_i = \pi_i = \frac{S}{K}$, that both Φ and Ψ are very incoherent ($\mu(\Phi), \mu(\Psi) \ll 1$) and well-conditioned ($\|\Psi\| \approx \|\Phi\| \approx \sqrt{K/d}$). Then condition (4.15) is equivalent to $\underline{\alpha}^2 \approx \frac{SK\log(K)}{Kd}$. By the relation $\varepsilon^2 = 2 - \underline{\alpha}$ this yields the sufficient condition

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

As the maximal distance between two dictionaries is $\sqrt{2}$, this is a huge improvement over existing results. The price we pay for this is that the cross-coherence $\mu(\Psi, \Phi)$ has to be very small in comparison to $\underline{\alpha}$. This is encapsulated by the following condition

$$\max_{i \neq j} |\langle \psi_i, \phi_j \rangle| \cdot \log(K) \lesssim \min_k |\langle \psi_k, \phi_k \rangle|.$$

Intuitively, one can think of this as ensuring that no two estimated atoms point to the same generating atom which might lead to errors in the sparse approximation algorithm. So if it is clear in some sense, which estimated atom belongs to which generating atom, then the admissible distance can be very close to $\sqrt{2}$.
Another set of conditions arise, if one has no information about the cross-coherence $\mu(\Psi, \Phi)$ but access to an incoherent and well-conditioned initialisation and generating dictionary, meaning

$$\max\left\{ \|\Phi D_{\sqrt{\pi}}\|^2, \|\Psi D_{\sqrt{\pi}}\|^2, \mu(\Phi), \mu(\Psi) \right\} \lesssim \frac{1}{\log(K)}.$$

In this setting we see that since $\mu(\Psi,\Phi)^2 \lesssim \mu(\Phi)^2+\varepsilon^2$, the condition $\varepsilon \lesssim 1/\sqrt{\log(K)}$ is sufficient for (4.15) to be satisfied. So as long as the generating dictionary and the initialisation are nice in some sense, the radius of convergence is $1/\sqrt{\log(K)}$ as stated in the introduction.
If one has no information at all about the structure or well-behavedness of the initialisation Ψ (via the coherence or cross-coherence), then one can always bound $\|\Psi D_{\sqrt{\pi}}\| \leq \|\Phi D_{\sqrt{\pi}}\|+\|ZD_{\sqrt{\pi}}\|$ and $\mu(\Psi) \lesssim \mu(\Phi)+\varepsilon$. Thus we see that as long as the generating dictionary is well-behaved, we have that $\delta = \max\{\|ZD_{\sqrt{\pi}}\|,\varepsilon\} \lesssim \frac{1}{\log(K)}$ is sufficient to ensure convergence. In the case of MOD even this last regime is a large improvement over the convergence radius $\varepsilon \lesssim 1/S^2$ derived in [5].

Attainable accuracy: Analysing the above conditions and assuming access to an arbitrarily close initialisation, i.e. $\Psi \approx \Phi$ and arbitrarily many signals N, we see that the attainable accuracy $\delta_\star$ depends on Φ via condition (4.15) which in approximation reduces to

$$\delta_\star \approx SK\gamma^{-2}\max_i \pi_i^{-2/3}\exp\left(-\frac{\underline{\alpha}^2}{\max\{\|\Phi D_{\sqrt{\pi}}\|^2,\mu(\Phi)\}}\right).$$

This shows nicely how the precision one can attain depends inversely on the frequency of the atom that appears rarest and on the properties of the generating dictionary via the weighted operator norm and coherence. So even with unlimited fresh samples in each iteration and an arbitrarily close initialisation, if the generating dictionary is not 'well-behaved', we might be limited in the accuracy up to which we can recover it.

Number of signals: From the probability bound in (4.16) we see that in order for the failure probability in each step to be small, the number of the fresh signals per iteration has to be approximately

$$N \approx \frac{\rho^2}{\delta_\star^2}\cdot\log(K) = \frac{4\kappa^4\|\Phi\|^4 S^2\gamma^{-4}\underline{\alpha}^{-4}\underline{\pi}^{-3}}{\delta_\star^2}\cdot\log(K)$$

ensuring that even the most rarely appearing atoms are seen often enough to learn them properly. For uniformly distributed supports where every atom is equally likely to be in the support, i.e. $\pi_i = S/K$, this means that N should be of order $K^3\log(K)/\delta_\star^2$. We are quite convinced that this bound can be reduced using more sophisticated variance bounds in the matrix and vector Bernstein inequalities later in the proof, however, we leave the endeavour to those still motivated after reading the current proof.

Conditioning of submatrices: The conditions

$$\max\left\{\|\Phi D_{\sqrt{\pi}}\|^2,\mu(\Phi)\right\} \lesssim \frac{1}{\log(K)} \quad \text{and} \quad \max\left\{\|\Psi D_{\sqrt{\pi}}\|^2,\mu(\Psi)\right\} \lesssim \frac{1}{\log(K)}$$

are quite standard assumptions in the theory of sparse approximation and dictionary learning since they ensure that **most** submatrices Φ_I and Ψ_I are well-behaved in the sense that $\|\Phi_I^*\Phi_I - \mathbb{I}\| \leq \vartheta < 1$ for most supports I. This allows us to work with the pseudo-inverse $\Phi_I^\dagger$ without worrying too much about too small or large singular values. Checking that this condition is still satisfied for the updated dictionary after each step is one of the main hurdles of the proof and the reason why we need to control also the weighted operator norm of the difference between Φ and the updated dictionary $\hat{\Psi}$.

4.5. Proof

To prove our main result we proceed as follows. We show that under the conditions of our main Theorem we have contraction in each iteration and that the conditions for this contraction are then again satisfied in the next step. For that recall that we defined the distance between the generating dictionary Φ and a guess Ψ as

$$\delta(\Psi, \Phi) := \max\left\{\|(\Psi - \Phi)D_{\sqrt{\pi}}\|_{2,2}, \|\Psi - \Phi\|_{2,1}\right\}.$$

Proposition 4.3 *Assume our signals follow the signal model in 4.1 with probabilities $p_1, \cdots, p_K$ such that $\sum_i p_i = S$ and $0 \leq p_i \leq \frac{1}{6}$. Let $\pi_i := \mathbb{P}_S(i \in I)$ and define*

$$\underline{\alpha} := \min_k |\langle \psi_k, \phi_k \rangle| = 1 - \varepsilon^2/2 \quad \text{and} \quad \gamma := \frac{c_{\min}}{c_{\max}} \quad \text{and} \quad \rho := 2\kappa^2 \|\Phi\|^2 S \gamma^{-2} \underline{\alpha}^{-2} \underline{\pi}^{-3/2}.$$

We denote by $\delta_\star$ the desired recovery accuracy and assume $\delta_\star \log(K\rho/\delta_\star) \leq \gamma^2/(8C)$. If the generating dictionary Φ satisfies

$$\max\left\{\|\Phi D_{\sqrt{\pi}}\|^2, \mu(\Phi)\right\} \leq \frac{1}{8C^2} \frac{\underline{\alpha}^2 \gamma^2}{\log(K\rho/\delta_\star)} \tag{4.17}$$

and the current guess Ψ satisfies either

$$\max\left\{\|\Psi D_{\sqrt{\pi}}\|^2, \mu(\Psi, \Phi)^2, \mu(\Psi)\right\} \leq \frac{1}{8C^2} \frac{\underline{\alpha}^2 \gamma^2}{\log(K\rho/\delta_\star)} \tag{4.18}$$

or

$$\delta(\Psi, \Phi) \leq \frac{1}{8C} \frac{\gamma^2}{\log(K\rho/\delta_\star)}, \tag{4.19}$$

for a universal constant $C > 1$, then the updated and normalised dictionary $\hat{\Psi}$ satisfies

$$\delta(\hat{\Psi}, \Phi) \leq \frac{1}{2}\delta_\star + \frac{1}{2} \min\left\{\frac{\gamma^2}{8C\log(K\rho/\delta_\star)}, \frac{\gamma}{8\sqrt{\log(K\rho/\delta_\star)}} \cdot \delta(\Psi, \Phi)\right\}, \tag{4.20}$$

except with probability

$$60K \exp\left(-\frac{N(\Delta/2)^2}{2\rho^2 + \rho\Delta/2}\right). \tag{4.21}$$

where

$$\Delta := \frac{1}{16}\delta_\star + \min\left\{\frac{\gamma^2}{128C\log(K\rho/\delta_\star)}, \frac{\gamma}{128\sqrt{\log(K\rho/\delta_\star)}} \cdot \delta(\Psi, \Phi)\right\}. \tag{4.22}$$

Remark 4.4 *A few remarks are in order. In contrast to the main theorem we have two sets of sufficient conditions on the current guess Ψ to ensure contraction. We will call* (4.18) *the first regime and* (4.19) *the second regime. The above proposition states that in the first regime — in which the distance between the generating dictionary and the initialisation may be close to $\sqrt{2}$ — we contract by a factor $\approx 1/\log(K)$, allowing us to jump into the second regime. Once in the second regime,* (4.20) *implies $\delta(\hat{\Psi}, \Phi) \leq \eta \cdot \delta(\Psi, \Phi)$ for some $\eta < \frac{5}{8}$, as long as $\delta(\Psi, \Phi) \geq \delta_\star$, so we converge with geometric rate up to the minimal attainable distance $\delta_\star$.*

Proof We show that under the following 4 claims, both dictionary learning algorithms contract towards the generating dictionary (in the maximum column norm and the weighted operator norm). For that define

$$q := K \exp\left(-\frac{N(\Delta/2)^2}{2\rho^2 + \rho\Delta/2}\right).$$

Claim 1 *Under the conditions of the proposition, we have*

$$\|\Phi A (D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1} - \Phi D_{\sqrt{\pi}}\|_{2,2} \leq \underline{\alpha}\Delta, \tag{C1}$$

except with probability $2q$.

Claim 2 *Under the conditions of the proposition, we have*

$$\|(D_{\sqrt{\pi}\cdot\alpha})^{-1} B (D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1} - \mathbb{I}\|_{2,2} \leq \Delta, \tag{C2}$$

except with probability $2q$.

Claim 3 *Under the conditions of the proposition, we have for all* $\ell \in \{1, \cdots, K\}$

$$\|\Phi A (D_{\pi\cdot\alpha\cdot\beta})^{-1} e_\ell - \phi_\ell\|_2 \leq \underline{\alpha}\Delta, \tag{C3}$$

except with probability $28q$.

Claim 4 *Under the conditions of the proposition, we have for all* $\ell \in \{1, \cdots, K\}$

$$\frac{\gamma\underline{\alpha}}{C\sqrt{\log(K\rho/\delta_\star)}} \cdot \|\mathbb{I}_{\ell^c} (D_{\sqrt{\pi}\cdot\alpha})^{-1} B (D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1} e_\ell \pi_\ell^{-\frac{1}{2}}\|_2 \leq \Delta, \tag{C4}$$

except with probability $28q$.

We will show that a properly scaled version of the updated dictionary, which we will denote by $\bar{\Psi}$, contracts towards the generating dictionary under the above claims. Concretely

$$\|(\bar{\Psi} - \Phi) D_{\sqrt{\pi}}\| \leq 4\Delta \quad \text{and} \quad \|\bar{\Psi} - \Phi\|_{2,1} \leq 4\Delta, \tag{4.23}$$

So we have contraction towards the generating dictionary in the weighted operator norm and the maximum column norm simultaneously.

aK-SVD: Recall that for aK-SVD, the updated dictionary before normalisation can be written in a concise way as

$$\tilde{\Psi} = \Phi X \hat{X}^* - \Psi \hat{X} \hat{X}^* + \Psi \operatorname{diag}(\hat{X}\hat{X}^*) = \Phi A - \Psi B + \Psi \operatorname{diag}(B).$$

In order to show that one dictionary update step decreases the distance to the generating dictionary Φ, we introduce a scaled dictionary update which we denote by $\bar{\Psi}$

$$\bar{\Psi} := \tilde{\Psi}(D_{\pi\cdot\alpha\cdot\beta})^{-1} = [\Phi A - \Psi B + \Psi \operatorname{diag}(B)]\,(D_{\pi\cdot\alpha\cdot\beta})^{-1}, \tag{4.24}$$

i.e., we multiply the dictionary update step of the K-SVD algorithm with the diagonal matrix $(D_{\pi\cdot\alpha\cdot\beta})^{-1}$ to ensure that, on average, this matrix concentrates around Φ. This does not change

the underlying algorithm, since we have a normalisation step at the end of each iteration, which will be analysed at the end of this section.
We make the following decomposition

$$\begin{aligned}\bar{\Psi} &= \Phi A (D_{\pi\cdot\alpha\cdot\beta})^{-1} - \Psi(B - \operatorname{diag}(B))(D_{\pi\cdot\alpha\cdot\beta})^{-1} \\ &= \Phi A (D_{\pi\cdot\alpha\cdot\beta})^{-1} - \Psi D_{\sqrt{\pi}\cdot\alpha}\left[(D_{\sqrt{\pi}\cdot\alpha})^{-1} B (D_{\pi\cdot\alpha\cdot\beta})^{-1} - D_{\sqrt{\pi}}^{-1}\right] \\ &\qquad + \Psi D_{\sqrt{\pi}\cdot\alpha}\left[(D_{\sqrt{\pi}\cdot\alpha})^{-1} \operatorname{diag}(B)(D_{\pi\cdot\alpha\cdot\beta})^{-1} - D_{\sqrt{\pi}}^{-1}\right]. \end{aligned} \tag{4.25}$$

Recall from above that we want to have concentration in the weighted operator norm and the maximal ℓ_2-distance. We begin by showing concentration in the weighted operator norm under claims C1-C4 and the assumptions on $\|\Psi D_{\sqrt{\pi}}\|$ (or $\|\Phi D_{\sqrt{\pi}}\| + \|(\Psi - \Phi)D_{\sqrt{\pi}}\|$). With the above expression for the updated dictionary $\bar{\Psi}$ we can bound the operator norm of the difference $(\bar{\Psi} - \Phi)D_{\sqrt{\pi}}$ as

$$\|(\bar{\Psi} - \Phi)D_{\sqrt{\pi}}\| \leq \underbrace{\|\Phi A (D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1} - \Phi D_{\sqrt{\pi}}\|}_{\leq \underline{\alpha}\Delta \text{ (C1)}} + 2 \underbrace{\|\Psi D_{\sqrt{\pi}\cdot\alpha}\|}_{\leq \frac{\underline{\alpha}\gamma}{C\sqrt{\log(K\rho/\delta_\star)}}} \underbrace{\|(D_{\sqrt{\pi}\cdot\alpha})^{-1} B (D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1} - \mathbb{I}\|}_{\leq \Delta \text{ (C2)}} \leq 4\Delta,$$

meaning that the scaled dictionary update step contracts towards the generating dictionary in the weighted operator norm.
Next, we are going to show that for each atom the ℓ_2-distance also decreases with each iteration. As for the operator norm we use the scaled version of the updated dictionary and access the ℓ-th dictionary atom $\bar{\psi}_\ell$ simply by multiplying $\bar{\Psi}$ with the standard basis vector e_ℓ. This yields

$$\begin{aligned}\bar{\psi}_\ell = \bar{\Psi} e_\ell &= \Phi A (D_{\pi\cdot\alpha\cdot\beta})^{-1} e_\ell + \Psi(B - \operatorname{diag}(B))(D_{\pi\cdot\alpha\cdot\beta})^{-1} e_\ell \\ &= \Phi A (D_{\pi\cdot\alpha\cdot\beta})^{-1} e_\ell + \Psi D_{\sqrt{\pi}\cdot\alpha}\left[(D_{\sqrt{\pi}\cdot\alpha})^{-1}(B - \operatorname{diag}(B))(D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1}\right] e_\ell \pi_\ell^{-\frac{1}{2}} \\ &= \Phi A (D_{\pi\cdot\alpha\cdot\beta})^{-1} e_\ell + \Psi D_{\sqrt{\pi}\cdot\alpha} \cdot \mathbb{I}_{\ell^c} \cdot \left[(D_{\sqrt{\pi}\cdot\alpha})^{-1} B (D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1}\right] e_\ell \pi_\ell^{-\frac{1}{2}}. \end{aligned} \tag{4.26}$$

Again using this decomposition together with claims C3 and C4 and the assumptions on $\|\Psi D_{\sqrt{\pi}}\|$ resp. $\|\Phi D_{\sqrt{\pi}}\| + \|(\Psi - \Phi)D_{\sqrt{\pi}}\|$ we get

$$\|\bar{\psi}_\ell - \phi_\ell\| \leq \underbrace{\|\Phi A (D_{\pi\cdot\alpha\cdot\beta})^{-1} e_\ell - \phi_\ell\|}_{\leq \underline{\alpha}\Delta \text{ (C3)}} + \underbrace{\underbrace{\|\Psi D_{\sqrt{\pi}\cdot\alpha}\|}_{\leq \frac{\gamma\underline{\alpha}}{C\sqrt{\log(K\rho/\delta_\star)}}} \|\mathbb{I}_{\ell^c} (D_{\sqrt{\pi}\cdot\alpha})^{-1} B (D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1} e_\ell \pi_\ell^{-\frac{1}{2}}\|}_{\leq \Delta \text{ (C4)}} \leq 4\Delta.$$

So putting the above together we have shown that under Claims 1-4, we get that

$$\|(\bar{\Psi} - \Phi)D_{\sqrt{\pi}}\|_{2,2} \leq 4\Delta \quad \text{and} \quad \|\bar{\Psi} - \Phi\|_{2,1} \leq 4\Delta, \tag{4.27}$$

i.e., that we have contraction in the weighted operator norm and in the maximum column norm. This does not finish the proof since we have to also take into account the normalisation step. We postpone the analysis of the normalising step to after the analysis of the MOD algorithm, since it is the same for both algorithms.

MOD: Turning to the MOD algorithm we recall that, if the estimated coefficient matrix $\hat{X}$ has full row rank K or equivalently $\hat{X}\hat{X}^*$ has full rank, which is guaranteed by Claim 2, we can write the dictionary update step before normalisation as

$$\tilde{\Psi} = \Phi X \hat{X}^* \left(\hat{X}\hat{X}^*\right)^{-1} = \Phi A B^{-1}.$$

This dictionary update step — though conceptually very easy — is harder to analyse theoretically due to the inverse of the matrix B. Since at the end of each iteration the current dictionary is normalised such that each atom has norm one, instead of analysing the actual dictionary update step before normalisation we look at a scaled version. This does not change anything, as the normalisation step simply cancels out this scaling, but it makes life much easier when analysing the dictionary update step. So we show that

$$\bar{\Psi} := \Phi A B^{-1} D_\alpha \approx \Phi. \tag{4.28}$$

So as above we start by showing that the weighted operator norm of the difference $\bar{\Psi} - \Phi$ contracts under our 4 claims from above. We have to be very careful with the inverse matrix B^{-1}. We will show that this matrix concentrates around $D_{\pi\cdot\alpha^2\cdot\beta}$. Concretely, we will split $(\bar{\Phi} - \Phi)D_{\sqrt{\pi}}$ as follows

$$\begin{aligned}(\bar{\Psi} - \Phi)D_{\sqrt{\pi}} &= \Phi A B^{-1} D_{\sqrt{\pi}\cdot\alpha} - \Phi D_{\sqrt{\pi}} \\ &= \Phi A (D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1} - \Phi D_{\sqrt{\pi}} + \Phi A \left[B^{-1} - (D_{\pi\cdot\alpha^2\cdot\beta})^{-1}\right] D_{\sqrt{\pi}\cdot\alpha} \\ &= \Phi A (D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1} - \Phi D_{\sqrt{\pi}} + \Phi A (D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1} \left[\left[(D_{\sqrt{\pi}\cdot\alpha})^{-1} B (D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1}\right]^{-1} - \mathbb{I}\right].\end{aligned} \tag{4.29}$$

Before bounding the operator norm of the above terms we will have a closer look at the expression in the last bracket. For ease of notation set

$$C := (D_{\sqrt{\pi}\cdot\alpha})^{-1} B (D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1}. \tag{4.30}$$

To analyse the inverse in last bracket of (4.29) we will make use of the Neumann series expansion. For that note that by (C2) we have that $\|C - \mathbb{I}\| \leq 1/2$, hence we can apply the Neumann series expansion on the matrix C^{-1}. This amounts to

$$(C^{-1} - \mathbb{I}) = \sum_{k=0}^{\infty} (C - \mathbb{I})^k - \mathbb{I} = \sum_{k=1}^{\infty} (C - \mathbb{I})^k = \sum_{k=0}^{\infty} (C - \mathbb{I})^k (C - \mathbb{I}) \tag{4.31}$$

So by (C2) we get for the operator norm of the above

$$\|C^{-1} - \mathbb{I}\| = \|\sum_{k=0}^{\infty} (C - \mathbb{I})^k (C - \mathbb{I})\| \leq \sum_{k=0}^{\infty} \|C - \mathbb{I}\|^k \|C - \mathbb{I}\| \leq \sum_{k=0}^{\infty} 1/2^k \|C - \mathbb{I}\| \leq 2\|C - \mathbb{I}\|.$$

Note also that by (C1), we have

$$\begin{aligned}\|\Phi A (D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1}\| &\leq \|\Phi A (D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1} - \Phi D_{\sqrt{\pi}} + \Phi D_{\sqrt{\pi}}\| \\ &\leq \underline{\alpha}\Delta + \|\Phi D_{\sqrt{\pi}}\| \leq \frac{\underline{\alpha}\gamma}{4C\sqrt{\log(K\rho/\delta_\star)}}.\end{aligned} \tag{4.32}$$

Putting all of these observations back into (4.29) and using the triangle inequality repeatedly we get

$$\|(\bar{\Psi} - \Phi)D_{\sqrt{\pi}}\| \leq \underbrace{\|\Phi A (D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1} - \Phi D_{\sqrt{\pi}}\|}_{\leq \underline{\alpha}\Delta \text{ (C1)}} + \underbrace{\|\Phi A (D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1}\|}_{\leq \frac{\underline{\alpha}\gamma}{4C\sqrt{\log(K\rho/\delta_\star)}} \text{ (4.32)}} \cdot 2 \cdot \underbrace{\|C - \mathbb{I}\|}_{\leq \Delta \text{ (C2)}} \leq 3\Delta. \tag{4.33}$$

This shows that under the assumptions of the proposition, the weighted operator norm of the distance between the generating dictionary and the scaled update decreases in each iteration.

Now to the contraction of every atom in the ℓ_2-norm. We will show this contraction only for the first atom, as all the others are analogous. The next step makes use of the Schur-decomposition and the separation of the inverse into different parts. First we again split our updated dictionary atom $\bar{\psi}_\ell$ into two parts

$$\bar{\psi}_\ell - \phi_\ell = \bar{\Psi} e_\ell - \phi_\ell = \left[\Phi A (D_{\pi\cdot\alpha\cdot\beta})^{-1} e_\ell - \phi_\ell\right] + \Phi A (D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1} (C^{-1} - \mathbb{I}) e_\ell \pi_\ell^{-\frac{1}{2}}. \tag{4.34}$$

Here (as in the previous section) the first term is well-behaved and makes no problems. (C3) implies

$$\|\Phi A (D_{\pi\cdot\alpha\cdot\beta})^{-1} e_\ell - \phi_\ell\| \leq \Delta. \tag{4.35}$$

The second term of (4.34) needs more work, as we have to control the ℓ-th column of the matrix $C^{-1} - \mathbb{I}$. The trick is to see that we have to split the inverse C^{-1} into an off-diagonal term, which is small enough to control the term $1/\sqrt{\pi_\ell}$ and an on-diagonal term, which only uses the ℓ-th column of the matrix $A(D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1} \approx \Phi D_{\sqrt{\pi}}$, which is small enough to control the term $1/\sqrt{\pi_\ell}$. To formalise this argument, we write

$$\begin{aligned}
&\|\Phi A (D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1} (C^{-1} - \mathbb{I}) e_\ell \pi_\ell^{-\frac{1}{2}}\| \\
&\leq \|\Phi A (D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1} e_\ell e_\ell^* (C^{-1} - \mathbb{I}) e_\ell \pi_\ell^{-\frac{1}{2}}\| + \|\Phi A (D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1} \mathbb{I}_{\ell^c} C^{-1} e_\ell \pi_\ell^{-\frac{1}{2}}\| \\
&\leq \|\Phi A (D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1} e_\ell \pi_\ell^{-\frac{1}{2}}\| \cdot \|e_\ell^* \left[C^{-1} - \mathbb{I}\right] e_\ell\| + \|\Phi A (D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1} \mathbb{I}_{\ell^c}\| \cdot \|\mathbb{I}_{\ell^c} C^{-1} e_\ell \pi_\ell^{-\frac{1}{2}}\| \\
&\leq \|\Phi A (D_{\pi\cdot\alpha\cdot\beta})^{-1} e_\ell\| \cdot \|C^{-1} - \mathbb{I}\| + \|\Phi A (D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1}\| \cdot \|\mathbb{I}_{\ell^c} C^{-1} e_\ell\| \cdot \pi_\ell^{-\frac{1}{2}},
\end{aligned} \tag{4.36}$$

effectively splitting this operator norm into two parts which will be dealt with separately. The last norm term of the above, $\|\mathbb{I}_{\ell^c} C^{-1} e_\ell\|$ still needs some special treatment before we can effectively bound it with our claims. Taking a closer look we see that we have to bound the ℓ-th column — without the ℓ-th row — of the matrix C^{-1}. For this we will use Schur's formula for matrix inversion.

Lemma 4.5 *For a square matrix $C \in \mathbb{R}^{K\times K}$, if*

$$C = \begin{bmatrix} a & c^* \\ c & D \end{bmatrix} \quad \text{then} \quad C^{-1} = \begin{bmatrix} a & c^* \\ c & D \end{bmatrix}^{-1} = \begin{bmatrix} a^{-1} + a^{-1}c^* M c a^{-1} & -a^{-1}c^* M \\ -Mca^{-1} & M \end{bmatrix}, \tag{4.37}$$

where $M := (D - ca^{-1}c^*)^{-1}$ is the Schur complement of a in the above matrix.

In our case, after rearranging and remembering the restriction matrices R_I with $A_I = AR_I$, we have $a := e_\ell^* C e_\ell$, $c := R_{\ell^c}^* \cdot C \cdot e_\ell$ and $M := R_{\ell^c}^* \cdot C^{-1} \cdot R_{\ell^c}$ with $\|M\| \leq \|C^{-1}\|$. Since for any matrix V the vector $\mathbb{I}_{\ell^c} V e_\ell$ differs from $R_{\ell^c}^* V e_\ell$ only by an extra zero entry their norms coincide and Schur's decomposition lemma implies

$$\|\mathbb{I}_{\ell^c} C^{-1} e_\ell\| = \|R_{\ell^c}^* C^{-1} e_\ell\| = \|a^{-1} M c\| \leq \underbrace{\|(C_{\ell,\ell})^{-1}\|}_{\leq \frac{1}{1-\Delta}\ \text{(C2)}} \cdot \underbrace{\|R_{\ell^c}^* C^{-1} R_{\ell^c}\|}_{\leq 2\ \text{(C2), (4.31)}} \cdot \|R_{\ell^c}^* C e_\ell\| \leq 4 \|\mathbb{I}_{\ell^c} C e_\ell\|.$$

Also by (C3)

$$\|\Phi A (D_{\pi\cdot\alpha\cdot\beta})^{-1} e_\ell\| \leq 1 + \Delta \leq \frac{3}{2}. \tag{4.38}$$

Putting all these observations together

$$\begin{aligned}&\|\Phi A(D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1}(C^{-1}-\mathbb{I})e_\ell\pi_\ell^{-\frac{1}{2}}\|\\&\quad\leq \underbrace{\|\Phi A(D_{\pi\cdot\alpha\cdot\beta})^{-1}e_\ell\|}_{\leq\frac{3}{2}\ (4.38)}\underbrace{\|C^{-1}-\mathbb{I}\|}_{\leq\Delta\ (C2)}+\underbrace{\underbrace{4\|\Phi A(D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1}\|}_{\leq\frac{\gamma\alpha}{C\sqrt{\log(K\rho/\delta_\star)}}\ (4.32)}\|\mathbb{I}_{\ell^c}Ce_\ell\pi_\ell^{-\frac{1}{2}}\|}_{\leq\Delta\ (C4)}\leq 3\Delta,\end{aligned}\tag{4.39}$$

Plugging (4.35) and (4.39) into (4.34) finally yields the ℓ_2-norm bound on the distance between the ℓ-th atom of the generating dictionary and the scaled updated dictionary

$$\|\bar{\psi}_\ell-\phi_\ell\|\leq\|\Phi A(D_{\pi\cdot\alpha\cdot\beta})^{-1}e_\ell-\phi_\ell\|+\|\Phi A(D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1}(C^{-1}-\mathbb{I})e_\ell\pi_\ell^{-\frac{1}{2}}\|\leq 4\Delta.\tag{4.40}$$

Normalisation Combining the above results shows that with high probability, the dictionary update step of each algorithm (with scaling) satisfies

$$\delta(\bar{\Psi},\Phi)=\max\left\{\|(\bar{\Psi}-\Phi)D_{\sqrt{\pi}}\|_{2,2},\|\bar{\Psi}-\Phi\|_{2,1}\right\}\leq 4\Delta.\tag{4.41}$$

So what is left to show is that the normalisation step at the end of iteration does not interfere with convergence. Let $F:=\operatorname{diag}(\|\bar{\psi}_i\|_2)^{-1}$ be the square diagonal normalization matrix and denote by $\hat{\Psi}:=\bar{\Psi}F$ the normalized dictionary of the current update step. Since $\|\phi_i\|_2=1$ we have

$$\|F\|_{2,2}\leq\frac{1}{1-\varepsilon(\bar{\Psi},\Phi)}\quad\text{and}\quad\|\mathbb{I}-F\|\leq\frac{\varepsilon(\bar{\Psi},\Phi)}{1-\varepsilon(\bar{\Psi},\Phi)}.\tag{4.42}$$

Hence the weighted operator norm of the difference of the generating dictionary Φ and the normalised update $\hat{\Psi}$ can be bounded as

$$\begin{aligned}\|(\hat{\Psi}-\Phi)D_{\sqrt{\pi}}\|&=\|(\bar{\Psi}F-\Phi)D_{\sqrt{\pi}}\|\leq\|(\bar{\Psi}-\Phi)D_{\sqrt{\pi}}F+\Phi D_{\sqrt{\pi}}(\mathbb{I}-F)\|\\&\leq\|(\bar{\Psi}-\Phi)D_{\sqrt{\pi}}F\|+\|\Phi D_{\sqrt{\pi}}(\mathbb{I}-F)\|\\&\leq\|(\bar{\Psi}-\Phi)D_{\sqrt{\pi}}\|\|F\|+\|\Phi D_{\sqrt{\pi}}\|\|(\mathbb{I}-F)\|\\&\leq\|(\bar{\Psi}-\Phi)D_{\sqrt{\pi}}\|\frac{1}{1-\varepsilon(\bar{\Psi},\Phi)}+\|\Phi D_{\sqrt{\pi}}\|\frac{\varepsilon(\bar{\Psi},\Phi)}{1-\varepsilon(\bar{\Psi},\Phi)}\\&\leq\frac{1+\|\Phi D_{\sqrt{\pi}}\|}{1-\varepsilon(\bar{\Psi},\Phi)}\cdot\max\left\{\|(\bar{\Psi}-\Phi)D_{\sqrt{\pi}}\|,\|\bar{\Psi}-\Phi\|_{2,1}\right\}\leq 8\Delta.\end{aligned}\tag{4.43}$$

The ℓ_2-norm can be bounded in a similar fashion

$$\|\hat{\Psi}-\Phi\|_{2,1}=\|\bar{\Psi}F-\Phi\|_{2,1}\leq\|\bar{\Psi}-\Phi\|_{2,1}\frac{1}{1-\varepsilon(\bar{\Psi},\Phi)}+\frac{\varepsilon(\bar{\Psi},\Phi)}{1-\varepsilon(\bar{\Psi},\Phi)}\leq 8\Delta.\tag{4.44}$$

Thus we get by definition of Δ for the normalised dictionary $\hat{\Psi}$

$$\delta(\hat{\Psi},\Phi)\leq 8\Delta=\frac{1}{2}\delta_\star+\frac{1}{2}\min\left\{\frac{\gamma^2}{8C\log(K\rho/\delta_\star)},\frac{\gamma}{8\sqrt{\log(K\rho/\delta_\star)}}\cdot\delta(\Psi,\Phi)\right\}.$$

Combining the probability estimates of Claims 1-4 finishes the proof. ■

Proof [Theorem 4.2] To proof our main Theorem 4.2 we only have to show that we are able to repeatedly apply Proposition 4.3. But this follows immediately from the speed of convergence (4.20) and the conditions of regime 2 (4.19) together with the assumption of $\delta_\star \leq \frac{1}{8C}\frac{\gamma^2}{\log(K\rho/\delta_\star)}$. This ensures that δ converges to $\delta_\star$ geometrically and thus finishes the proof. ■

4.6. Technical results

Since we are going to apply the matrix and vector Bernstein inequalities repeatedly, we have to have good control over the operator norm of expectations of products of random matrices. A crucial technical result to do so is the following, which can be found in [27], [43].

Lemma 4.6 (Sum of random matrices [27], [43]) *Let $A_n \in \mathbb{R}^{d_1 \times d_2}$, $B_n \in \mathbb{R}^{d_2 \times d_3}$, $C_n \in \mathbb{R}^{d_3 \times d_4}$. Then*

$$\left\| \sum_{n=1}^{N} A_n B_n C_n \right\| \leq \left\| \sum_{n=1}^{N} A_n A_n^* \right\|^{1/2} \max_n \|B_n\| \left\| \sum_{n=1}^{N} C_n^* C_n \right\|^{1/2}.$$

Proof Write

$$\sum_{n=1}^{N} A_n B_n C_n = \begin{pmatrix} A_1 & A_2 & A_3 & . \\ . & . & & \\ . & & . & \\ . & & & . \end{pmatrix} \begin{pmatrix} B_1 & . & . & . \\ . & B_2 & & \\ . & & B_3 & \\ . & & & . \end{pmatrix} \begin{pmatrix} C_1 & . & . & . \\ C_2 & . & & \\ C_3 & & . & \\ . & & & . \end{pmatrix}. \tag{4.45}$$

Now the result immediately follows by applying the following properties of the operator norm $\|ABC\| \leq \|A\|\|B\|\|C\|$, $\|A\| = \|AA^*\|^{1/2}$ and $\|C\| = \|C^*C\|^{1/2}$. ■

This result immediately translates to expectations of products of random matrices with the following nice little trick.

Lemma 4.7 *[black magic box 2.0] Let $A(I) \in \mathbb{R}^{d_1 \times d_2}$, $B(I) \in \mathbb{R}^{d_2 \times d_3}$, $C(I) \in \mathbb{R}^{d_3 \times d_4}$ be random matrices, where I is a discrete random variable taking values in $\mathcal{I}$. Then for any $\mathcal{G} \subseteq \mathcal{I}$*

$$\|\mathbb{E}\left[A(I)B(I)C(I)\mathbb{1}_{\mathcal{G}}(I)\right]\| \leq \|\mathbb{E}\left[A(I)A(I)^*\right]\|^{1/2} \cdot \max_{I \in \mathcal{G}} \|B(I)\| \cdot \|\mathbb{E}\left[C(I)^*C(I)\right]\|^{1/2}$$

Proof Rewriting the expectation as a sum and applying the lemma above yields

$$\begin{aligned} \|\mathbb{E}\left[A(I)B(I)C(I)\mathbb{1}_{\mathcal{G}}(I)\right]\| &= \|\sum_{I \in \mathcal{G}} \mathbb{P}[I]^{1/2} A(I)B(I)C(I)\mathbb{P}[I]^{1/2}\| \\ &\leq \|\sum_{I \in \mathcal{G}} \mathbb{P}[I]A(I)A(I)^*\|^{1/2} \cdot \max_{I \in \mathcal{G}} \|B(I)\| \cdot \|\sum_{I \in \mathcal{G}} \mathbb{P}[I]C(I)^*C(I)\|^{1/2} \\ &\leq \|\mathbb{E}\left[A(I)A(I)^*\right]\|^{1/2} \cdot \max_{I \in \mathcal{G}} \|B(I)\| \cdot \|\mathbb{E}\left[C(I)^*C(I)\right]\|^{1/2}, \end{aligned}$$

where in the last inequality we have used that the matrices $A(I)A(I)^*$ and $C(I)^*C(I)$ are positive semidefinite and that $P[I] \geq 0$. ■

Since all of our proofs rely on either the vector or matrix Bernstein inequality or the Chernoff inequality, we recall them here for convenience.

Theorem 4.8 (Matrix resp. Vector Bernstein [84, 52]) *Consider a sequence $Y_1, ..., Y_N$ of independent, random matrices (resp. vectors) with dimension $d \times K$ (resp. d). Assume that each random matrix (resp. vector) satisfies*

$$\|Y_n\| \leq R \quad a.s. \quad and \quad \|\mathbb{E}[Y_n]\| \leq m.$$

Then, for all $t > 0$,

$$\mathbb{P}\left(\|\frac{1}{N}\sum_{n=1}^{N} Y_n\| \geq m + t\right) \leq \kappa \exp\left(\frac{-Nt^2}{2R^2 + (R+m)t}\right), \tag{4.46}$$

where $\kappa = d + K$ *for the matrix Bernstein inequality and* $\kappa = 28$ *for the vector Bernstein inequality.*

Theorem 4.9 (Matrix Chernoff Inequality [84]) *Let* $X_1, ..., X_N$ *be independent random positive semi-definite matrices taking values in* $\mathbb{R}^{d\times d}$. *Assume that for all* $n \in \{1, ..., N\}$, $\|X_n\| \leq \eta$ *a.s. and* $\|\sum_{n=1}^{N} \mathbb{E}[X_n]\| \leq \mu_{max}$. *Then, for all* $r \geq e\mu_{max}$,

$$\mathbb{P}\left(\|\sum_{n=1}^{N} X_n\| \geq r\right) \leq K\left(\frac{e\mu_{max}}{r}\right)^{\frac{r}{\eta}}.$$

4.7. Sparse approximation and conditioning of subdictionaries

A major hurdle in analysing the above dictionary learning algorithms is that each update of the sparse coefficients involves projecting onto submatrices of the current guess Ψ. For the remainder of this chapter we will write

$$\mathcal{F}_\Phi := \{I \ : \ \|\Phi_I^*\Phi_I - \mathbb{I}\| \leq \vartheta\} \quad \text{and} \quad \mathcal{F}_\Psi := \{I \ : \ \|\|\Psi_I^*\Psi_I - \mathbb{I}\| \leq \vartheta\}$$

for the set of index sets where the random variables Φ_I resp. Ψ_I are well conditioned. We further write

$$\mathcal{F}_Z := \left\{I \ : \ \|Z_I\| \leq \delta \cdot e\sqrt{2\log(320K\rho/\delta_\star)}\right\}.$$

for the set of index sets, where the norm of the random variable Z_I is comparable to δ. Finally, set

$$\mathcal{G} := \mathcal{F}_\Phi \cup \mathcal{F}_\Psi \cup \mathcal{F}_Z. \tag{4.47}$$

To control $\mathbb{P}(\mathcal{F}_\Phi^c)$ and $\mathbb{P}(\mathcal{F}_\Psi^c)$ we use Theorem 2.1 from Chapter 2 which we recall here for convenience.

Theorem 4.10 (Operator norm of a random submatrix 2.1) *Let Ψ be a dictionary and assume $I \subseteq \mathbb{K}$ is chosen according to the rejective sampling model with probabilities $p_1, \ldots, p_K$ such that $\sum_{i=1}^K p_i = S$. Recall that $\pi_i := \mathbb{P}_S(i \in I)$. Further let D_π denote the diagonal matrix with the vector π on its diagonal. Then*

$$\mathbb{P}\left(\|\Psi_I^*\Psi_I - \mathbb{I}\| > r\right) \leq 216K \exp\left(-\min\left\{\frac{r^2}{4e^2\|\Psi D_\pi \Psi^*\|}, \frac{r}{2\mu(\Psi)}\right\}\right).$$

Further we need to control the sparse approximation step in each iteration. Recall that thresholding works by finding the indices corresponding to the S largest values of $|\langle\psi_i, y\rangle|$, i.e.

$$\begin{aligned} &\text{find} \quad \hat{I} \in \operatorname{argmax}_{|I|=S} \|\Psi_I^* y\|_1 \quad \text{and} \\ &\text{reconstruct} \quad \hat{x}_{\hat{I}} = \Psi_{\hat{I}}^\dagger y. \end{aligned}$$

In [74], average case results for thresholding were derived for the uniform case. There, a sufficient condition for thresholding to work with high probability was $S\mu^2\log(K) \lesssim \gamma^2$. The recent work [67] extended upon these results. Note that these results only apply for the case where one has knowledge of the generating dictionary Φ which is not the case in dictionary learning. This setting, where knowledge of the generating dictionary is not give, was covered in the recent work [57]. We adapt their result to our setting. For the remainder of this chapter, we write

$$\mathcal{H} := \left\{(I, \sigma, c) \mid \hat{I} = I\right\} \tag{4.48}$$

for the set of index, sign and coefficient triplets, where thresholding is guaranteed to recover the correct support. With all the necessary notation in place, we finally show that under the assumptions of our Proposition 4.3, the failure probability of Thresholding and the probability that our submatrices are ill-conditioned can be bound by approximately $\delta_\star/\rho$. This will be used repeatedly by the Lemmas thereafter.

Lemma 4.11 *Under the assumptions of Proposition 4.3 we have*

$$\mathbb{P}(\mathcal{H}^c) \cdot 2\rho + \mathbb{P}(\mathcal{G}^c) \cdot \rho \leq \frac{1}{32}\delta_\star. \tag{4.49}$$

Proof We begin with bounding the failure probability of Thresholding, $\mathbb{P}(\mathcal{H}^c)$. Set $\mathit{M} := \Psi^*\Phi - \operatorname{diag}(\Psi^*\Phi)$. By definition of the algorithm, thresholding recovers the full support of a signal $y = \Phi_I x_I$, if

$$\|\Psi_{I^c}^* y\|_\infty < \|\Psi_I^* y\|_{\min}.$$

Note that the signals have two sources of randomness, σ and I. Recall $\underline{\alpha} = \min_i |\langle \psi_i, \phi_i\rangle|$. Plugging in the definition of y we derive a bound on the failure probability

$$\begin{aligned}\mathbb{P}_y(\|\Psi_I^* y\|_{\min} < \|\Psi_{I^c}^* y\|_\infty) &= \mathbb{P}_y\left(\|\Psi_I^*\Phi_I x_I\|_{\min} < \|\Psi_{I^c}^*\Phi_I x_I\|_\infty\right)\\ &\le \mathbb{P}_y\left(c_{\min}\|\operatorname{diag}(\Psi_I^*\Phi_I)\|_{\min} - \|(\Psi_I^*\Phi_I - \operatorname{diag}(\Psi_I^*\Phi_I))x_I\|_\infty < \|\Psi_{I^c}^*\Phi_I x_I\|_\infty\right)\\ &\le \mathbb{P}_y\left(c_{\min}\cdot\underline{\alpha} < 2\|\mathit{M}_I x_I\|_\infty\right). \end{aligned} \tag{4.50}$$

Next we use that for $k \in I$, we have $x_k = \sigma_k c_k$, where $\sigma \in \mathbb{R}^S$ is an independent Rademacher sequence. Now as the signs σ are independent from the support I, we can apply Hoeffding's inequality to each entry of $\mathit{M}_I x_I$ to get

$$\begin{aligned}&\mathbb{P}_y(\|\Psi_I^* y\|_{\min} < \|\Psi_{I^c}^* y\|_\infty)\\ &\quad\le \mathbb{P}_y\left(\|\mathit{M}_I x_I\|_\infty \ge \frac{c_{\min}}{2}\cdot\underline{\alpha}\;\middle|\;\|\mathit{M}_I\|_{\infty,2} < \eta\right) + \mathbb{P}_S\left(\|\mathit{M}_I\|_{\infty,2} \ge \eta\right)\\ &\quad\le 2K\exp\left(-\frac{c_{\min}^2}{8c_{\max}^2\eta^2}\cdot\underline{\alpha}^2\right) + 2K\left(e\frac{\|\mathit{M}D_{\sqrt{p}}\|_{\infty,2}^2}{\eta^2}\right)^{\frac{\eta^2}{\mu(\Psi,\Phi)^2}}\\ &\quad\le 2K\exp\left(-\frac{\gamma^2}{8\eta^2}\cdot\underline{\alpha}^2\right) + 2K\left(2e\frac{\|\mathit{M}D_{\sqrt{\pi}}\|_{\infty,2}^2}{\eta^2}\right)^{\frac{\eta^2}{\mu(\Psi,\Phi)^2}}, \end{aligned} \tag{4.51}$$

where we used that by Theorem 4.17(a) we have $p_k \le 2\pi_k$ which implies $\|\mathit{M}D_{\sqrt{p}}\|_{\infty,2}^2 \le 2\|\mathit{M}D_{\sqrt{\pi}}\|_{\infty,2}^2$. Further we can bound $\|\mathit{M}D_{\sqrt{\pi}}\|_{\infty,2}^2$ as

$$\|\mathit{M}D_{\sqrt{\pi}}\|_{\infty,2}^2 = \|(\Psi^*\Phi - \operatorname{diag}(\Psi^*\Phi))D_{\sqrt{\pi}}\|_{\infty,2}^2 \le \|\Phi D_{\sqrt{\pi}}\|_{2,2}^2 = \|\Phi D_\pi \Phi^*\|_{2,2} \tag{4.52}$$

and thus by setting $\eta^2 := \frac{\gamma^2}{8\log(4K/\varepsilon)}\cdot\underline{\alpha}^2$ we get

$$\mathbb{P}\left(\mathcal{H}^c\right) \le 4K\exp\left(-\min\left\{\frac{\gamma^2\underline{\alpha}^2}{16e^2\|\Phi D_{\sqrt{\pi}}\|_{2,2}^2}, \frac{\gamma^2\underline{\alpha}^2}{8\mu(\Phi,\Psi)^2}\right\}\right). \tag{4.53}$$

Now we turn to bounding the quanitity $\mathbb{P}(\mathcal{F}_Z^c)$. Setting $t = 2e^2\delta^2\log(320K\rho/\delta_\star)$ we have by the Poissonisation trick 2, Lemma 2.5

$$\mathbb{P}_S(\mathcal{F}_Z^c) = \mathbb{P}_S\big(\|Z_I\|^2 > t\big) = \mathbb{P}_S\big(\|Z_I Z_I^*\| > t\big) \le 2\mathbb{P}_B\big(\|Z_I Z_I^*\| > t\big),$$

where $\mathbb{P}_B$ is the Poisson sampling model corresponding to the $p_1, \cdots, p_K$. Now a simple application of the Matrix Chernoff Inequality 4.9 together with 4.17(a) in the last inequality yields

$$\begin{aligned}\mathbb{P}_S\big(\|Z_I Z_I^*\| > t\big) &\le 2\mathbb{P}_B\big(\|Z_I Z_I^*\| > t\big)\\ &\le 2K\left(\frac{e\|ZD_pZ^*\|}{t}\right)^{t/\varepsilon^2} \le 2K\left(\frac{2e\|ZD_\pi Z^*\|}{t}\right)^{t/\varepsilon^2} \le 2K\left(\frac{2e\delta^2}{t}\right)^{t/\delta^2}\end{aligned}$$

Plugging in $t = 2e^2\delta^2 \log(320K\rho/\delta_\star)$ we get

$$\mathbb{P}_S(\mathcal{F}_Z^c) = \mathbb{P}_S\left(\|Z_I\|^2 \geq 2e^2\delta^2 \log(320K\rho/\delta_\star)\right) \leq \frac{\delta_\star}{160\rho}. \tag{4.54}$$

Applying Theorem 4.10 to to bound $\mathbb{P}(\mathcal{F}_\Phi^c)$ and $\mathbb{P}(\mathcal{F}_\Psi^c)$ and collecting all constants into universal constants m_1 to m_3 yields thus

$$\begin{aligned}
\mathbb{P}(\mathcal{H}^c) &\leq m_1 K \exp\left(-\min\left\{\frac{\gamma^2\underline{\alpha}^2}{\|\Phi D_{\sqrt{\pi}}\|^2}, \frac{\gamma^2\underline{\alpha}^2}{\mu(\Phi,\Psi)^2}\right\}\right) \\
\mathbb{P}(\mathcal{G}^c) &\leq m_2 K \exp\left(-\min\left\{\frac{1}{\|\Psi D_{\sqrt{\pi}}\|^2}, \frac{1}{\mu(\Psi)}\right\}\right) \\
&\quad + m_3 K \exp\left(-\min\left\{\frac{1}{\|\Phi D_{\sqrt{\pi}}\|^2}, \frac{1}{\mu(\Phi)}\right\}\right) + \frac{\delta_\star}{160\rho}.
\end{aligned} \tag{4.55}$$

So if we have for a universal constant $C > 0$ that

$$\max\left\{\|\Phi D_{\sqrt{\pi}}\|^2, \|\Psi D_{\sqrt{\pi}}\|^2, \mu(\Phi), \mu(\Psi,\Phi)^2, \mu(\Psi)\right\} \leq \frac{1}{C}\frac{\underline{\alpha}^2\gamma^2}{\log\left(K\rho/\delta_\star\right)}, \tag{4.56}$$

then the claim follows. In the first regime of the Theorem, this follows immediately from the assumptions on the generating dictionary Φ and the current guess Ψ. In the second regime we again see that the conditions on $\|\Phi D_{\sqrt{\pi}}\|$ and $\mu(\Phi)$ are satisfied by assumption. What is left to show is that $\max\{\mu(\Psi,\Phi)^2, \mu(\Psi), \|\Psi D_{\sqrt{\pi}}\|^2\} \leq \frac{1}{C}\frac{\underline{\alpha}^2\gamma^2}{\log(K\rho/\delta_\star)}$. For this recall that the assumption reads as

$$\delta(\Psi,\Phi) = \max\left\{\|(\Psi-\Phi)D_{\sqrt{\pi}}\|, \|\Psi-\Phi\|_{2,1}\right\} \leq \frac{1}{8C}\frac{\gamma^2}{\log\left(K\rho/\delta_\star\right)}. \tag{4.57}$$

So splitting Ψ into Φ and $\Psi - \Phi$ yields

$$\|\Psi D_{\sqrt{\pi}}\|^2 \leq 2\|\Phi D_{\sqrt{\pi}}\|^2 + 2\|(\Psi-\Phi)D_{\sqrt{\pi}}\|^2 \leq \frac{1}{2C}\frac{\gamma^2}{\log\left(K\rho/\delta_\star\right)} \leq \frac{1}{C}\frac{\underline{\alpha}^2\gamma^2}{\log\left(K\rho/\delta_\star\right)}, \tag{4.58}$$

where the last inequality follows from $\underline{\alpha}^2 = (1-\varepsilon^2/2)^2 \geq \frac{1}{2}$ since we are in the regime $\varepsilon \lesssim \frac{1}{\log(K)}$. Further the coherence $\mu(\Psi)$ can be decomposed as

$$\begin{aligned}
\mu(\Psi) = \max_{i\neq j}|\langle\psi_i,\psi_j\rangle| &\leq \max_{i\neq j}|\langle\phi_i,\phi_j\rangle| + 2\varepsilon + \varepsilon^2 \\
&\leq \mu(\Phi) + 3\varepsilon \leq \frac{1}{2C}\frac{\gamma^2}{\log\left(K\rho/\delta_\star\right)} \leq \frac{1}{C}\frac{\underline{\alpha}^2\gamma^2}{\log\left(K\rho/\delta_\star\right)}.
\end{aligned} \tag{4.59}$$

Since $\mu(\Psi,\Phi)^2$ can in the same manner be decomposed into $\mu(\Phi)^2$ and ε^2, the assumptions (4.56) are also satisfied in the second regime and the claim follows. ∎

With this result we are finally able to proof Lemmas 4.12- 4.15 — corresponding to Claims 1-4.

4.8. Proof of Claims 1-4

Lemma 4.12 (proof of Claim 1) *Under the assumptions of Proposition 4.3 we have*

$$\mathbb{P}\left(\|\Phi A(D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1} - \Phi D_{\sqrt{\pi}}\| > \underline{\alpha}\Delta\right) \leq (d+K)\exp\left(-\frac{N(\Delta/2)^2}{2\rho^2 + \rho\Delta/2}\right), \tag{4.60}$$

where

$$\rho = 2\kappa^2\|\Phi\|^2 S\gamma^{-2}\underline{\alpha}^{-2}\underline{\pi}^{-3/2}.$$

Proof The idea is to write $\Phi A(D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1} - \Phi D_{\sqrt{\pi}}$ as a sum of independent random matrices and apply matrix Bernstein to show that we indeed have concentration. Since we assumed in the algorithm that the estimated coefficients can never have larger norm than the signal times κ we first define for $v \in \mathbb{R}^d$ the set of possible stable supports as $\mathcal{B}(v) := \{I : \|\Psi_I^\dagger v\| \leq \kappa\|v\|\}$. Based on this definition we further define the following random matrices for $n \in [N]$

$$\hat{Y}_n := y_n y_n^* (\Psi_{\hat{I}_n}^\dagger)^* R_{\hat{I}_n}(D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1} \cdot \mathbb{1}_{\mathcal{B}(y_n)}(\hat{I}_n) - \Phi D_{\sqrt{\pi}},$$

where as always, $\hat{I}_n$ denotes the set found by the thresholding algorithm. As each matrix $\hat{Y}_n$ only depends on the signal y_n they are independent and we have

$$\frac{1}{N}\sum_n \hat{Y}_n = \Phi A(D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1} - \Phi D_{\sqrt{\pi}},$$

so we can use the matrix Bernstein inequality to bound the left hand side of (4.12). For that we have to find an upper bound for the operator norm. Since we assumed in the algorithm that the estimated coefficients can never have larger norm than the signal - i.e. $\|\hat{x}_n\| = \|\Psi_{\hat{I}_n}^\dagger y_n\|_2 \leq \kappa\|y_n\|_2$ and $\|y_n\| \leq \|\Phi\|$ we get

$$\|\hat{Y}_n\| \leq \kappa\|\Phi\|^2 S c_{\max}^2 \|D_{\sqrt{\pi}}^{-1}\|\|D_\beta^{-1}\|\|D_\alpha^{-1}\| + \|\Phi D_{\sqrt{\pi}}\| \leq \frac{3}{4}\rho\underline{\alpha} =: R. \tag{4.61}$$

Bounding $\|\mathbb{E}[\hat{Y}_n]\|$ for some n is a little more involved. Recall that $\mathcal{H}$ is the set of signals y, meaning support, sign and coefficient triplets (I, σ, c), where thresholding recovers the correct support from the corresponding signal. Further $\mathcal{G}$ is the set of supports I where ϑ_I is small - i.e. the corresponding subdictionary Ψ_I is well-conditioned. Therefore for each n we define a new random matrix Y_n, for which the estimated support $\hat{I}_n$ is replaced with the correct support I_n and $\Phi D_{\sqrt{\pi}}$ is replaced by $\Phi\,\mathrm{diag}(\mathbf{1}_{I_n})D_{\sqrt{\pi}}^{-1}$.

$$Y_n := y_n y_n^* (\Psi_{I_n}^\dagger)^* R_{I_n}(D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1} \cdot \mathbb{1}_{\mathcal{B}(y_n)}(I_n) - \Phi\,\mathrm{diag}(\mathbf{1}_{I_n})D_{\sqrt{\pi}}^{-1}$$

Note that with the same argument as above Y_n is bounded by R. Further, by definition of $\mathcal{H}$ the first terms of the two random matrices coincide on $\mathcal{H}$, while the second terms coincide in expectation, meaning $\mathbb{E}[\Phi\,\mathrm{diag}(\mathbf{1}_I)D_{\sqrt{\pi}}^{-1}] = \Phi D_{\sqrt{\pi}}$. So dropping the index n for convenience, as each signal has the same distribution, e.g., writing I for I_n, we get

$$\begin{aligned}
\|\mathbb{E}[\hat{Y}]\| &\leq \|\mathbb{E}[\hat{Y} - Y]\| + \|\mathbb{E}[Y]\| \\
&\leq \mathbb{P}(\mathcal{H}^c)\cdot 2\rho\underline{\alpha} + \|\mathbb{E}[\mathbb{1}_{\mathcal{G}^c}(I)Y]\| + \|\mathbb{E}[\mathbb{1}_{\mathcal{G}}(I)Y]\| \\
&\leq \mathbb{P}(\mathcal{H}^c)\cdot 2\rho\underline{\alpha} + \mathbb{P}(\mathcal{G}^c)\cdot\rho\underline{\alpha} + \|\mathbb{E}[\mathbb{1}_{\mathcal{G}}(I)Y]\|
\end{aligned} \tag{4.62}$$

Next note that whenever $I \in \mathcal{G}$ we have for any sign and coefficient pair (σ, c) that the corresponding signal y satisfies $\|\Psi_I^\dagger y\| \leq (1-\vartheta_I)^{-\frac{1}{2}} \cdot \|y\| \leq \kappa\|y\|$, so we have $\mathcal{G} \subseteq B(y)$. Remembering that $y = \Phi_I x_I = \Phi_I(\sigma_I \odot c_I)$, we can therefore take the expectation over (σ, c). Using the shorthand $\mathbb{E}_{\mathcal{G}}[f(I)] := \mathbb{E}_I[\mathbb{1}_{\mathcal{G}}(I)f(I)]$ this yields

$$\begin{aligned}
\|\mathbb{E}[\mathbb{1}_{\mathcal{G}}(I)Y]\| &= \|\mathbb{E}_I[\mathbb{1}_{\mathcal{G}}(I)\cdot\mathbb{E}_{\sigma,c}[\Phi_I x_I x_I^* \Phi_I^* \Psi_I^{\dagger *} R_I(D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1} - \Phi\,\mathrm{diag}(\mathbf{1}_I)D_{\sqrt{\pi}}^{-1}]]\| \\
&= \|\mathbb{E}_{\mathcal{G}}[\Phi_I\Phi_I^*\Psi_I^{\dagger *} R_I(D_{\sqrt{\pi}\cdot\alpha})^{-1} - \Phi\,\mathrm{diag}(\mathbf{1}_I)D_{\sqrt{\pi}}^{-1}]\| \\
&\leq \|\Phi D_{\sqrt{\pi}}\|\cdot\|\mathbb{E}_{\mathcal{G}}[D_{\sqrt{\pi}}^{-1}R_I^*\Phi_I^*\Psi_I^{\dagger *}R_I(D_{\sqrt{\pi}\cdot\alpha})^{-1} - D_{\sqrt{\pi}}^{-1}\,\mathrm{diag}(\mathbf{1}_I)D_\alpha(D_{\sqrt{\pi}\cdot\alpha})^{-1}]\| \\
&\leq \|D_\alpha^{-1}\|\cdot\|\Phi D_{\sqrt{\pi}}\|\cdot\left\|\mathbb{E}_{\mathcal{G}}\left[D_{\sqrt{\pi}}^{-1}R_I^*\left(\Phi_I^*\Psi_I^{\dagger *} - (D_\alpha)_{I,I}\right)R_I D_{\sqrt{\pi}}^{-1}\right]\right\|.
\end{aligned} \tag{4.63}$$

To simplify further note that for $I \in \mathcal{G}$ we can write

$$\Psi_I^\dagger = (\Psi_I^*\Psi_I)^{-1}\Psi_I^* = (\Psi_I^*\Psi_I - \mathbb{I}_S + \mathbb{I}_S)^{-1}\Psi_I^* = \sum_{k=0}^{\infty}(\Psi_I^*\Psi_I - \mathbb{I}_S)^k\Psi_I^* = \sum_{k=0}^{\infty} H_{I,I}^k\Psi_I^*, \quad (4.64)$$

and have $\max_{I\in\mathcal{G}} \| \sum_{k=0}^{\infty} H_{I,I}^k\| \leq \frac{1}{1-\vartheta}$. Also since $Z = \Psi - \Phi$ we have by definition of D_α

$$(D_\alpha)_{I,I} = \mathbb{I}_S - \operatorname{diag}(\Psi_I^* Z_I). \quad (4.65)$$

Further, on $\mathcal{G}$ the columns of Ψ_I are linearly independent, meaning we have $\Psi_I^\dagger\Psi_I = \mathbb{I}_S$ and hence we get for the expression inside the expectation above

$$\begin{aligned}\Phi_I^*(\Psi_I^\dagger)^* - (D_\alpha)_{I,I} &= (\Psi_I^* - Z_I^*)(\Psi_I^\dagger)^* - (D_\alpha)_{I,I} \\ &= \mathbb{I}_S - Z_I^*\Psi_I^{\dagger *} - (D_\alpha)_{I,I} = Z_I^*\Psi_I\sum_{k=0}^{\infty} H_{I,I}^k - \operatorname{diag}(Z_I^*\Psi_I). \end{aligned} \quad (4.66)$$

Further we will use the decomposition $И_{I,I} := Z_I^*\Psi_I - \operatorname{diag}(Z_I^*\Psi_I) = Z_I^*\Psi_I - \mathcal{E}_{I,I}$. Plugging this back into the expression (4.63) we see that we have to control the operator norms of the expectations of the terms in (4.66). By using that $\|\Psi D_{\sqrt{\pi}}\| \leq \frac{1}{8}$, Corollary 4.20(c) implies the following inequalities which will be used multiple times

$$\|\mathbb{E}[D_{\sqrt{\pi}}^{-1}R_I^*H_{I,I}H_{I,I}^*R_ID_{\sqrt{\pi}}^{-1}]\| \leq \frac{9}{2}\|D_{\sqrt{\pi}}HD_{\sqrt{\pi}}\|^2 + \frac{3}{2}\|\Psi D_{\sqrt{\pi}}\|^2 \leq 2\|\Psi D_{\sqrt{\pi}}\|^2 \quad (4.67)$$

$$\|\mathbb{E}[D_{\sqrt{\pi}}^{-1}R_I^*И_{I,I}И_{I,I}^*R_ID_{\sqrt{\pi}}^{-1}]\| \leq 18\|\Psi D_{\sqrt{\pi}}\|^2\|ZD_{\sqrt{\pi}}\|^2 + \frac{3}{2}\varepsilon^2\|\Psi D_{\sqrt{\pi}}\|^2 \quad (4.68)$$

$$\|\mathbb{E}[D_{\sqrt{\pi}}^{-1}R_I^*И_{I,I}^*И_{I,I}R_ID_{\sqrt{\pi}}^{-1}]\| \leq 18\|\Psi D_{\sqrt{\pi}}\|^2\|ZD_{\sqrt{\pi}}\|^2 + \frac{3}{2}\|ZD_{\sqrt{\pi}}\|^2 \leq 2\|ZD_{\sqrt{\pi}}\|^2. \quad (4.69)$$

Further recall that $\vartheta \leq 1/4$ and $\varepsilon \leq \sqrt{2}$. With this we get

$$\begin{aligned}&\|\mathbb{E}_{\mathcal{G}}[D_{\sqrt{\pi}}^{-1}R_I^*\Big(Z_I^*\Psi_I\sum_{k=0}^{\infty}H_{I,I}^k - \mathcal{E}_{I,I}\Big)R_ID_{\sqrt{\pi}}^{-1}]\| \\ &= \|\mathbb{E}_{\mathcal{G}}[D_{\sqrt{\pi}}^{-1}R_I^*\Big(И_{I,I}\sum_{k=0}^{\infty}H_{I,I}^k + \mathcal{E}_{I,I}\sum_{k=1}^{\infty}H_{I,I}^k\Big)R_ID_{\sqrt{\pi}}^{-1}]\| \\ &\leq \|\mathbb{E}_{\mathcal{G}}[D_{\sqrt{\pi}}^{-1}R_I^*И_{I,I}\cdot\sum_{k=0}^{\infty}H_{I,I}^k\cdot R_ID_{\sqrt{\pi}}^{-1}]\| + \|\mathbb{E}_{\mathcal{G}}[D_{\sqrt{\pi}}^{-1}R_I^*H_{I,I}\cdot\sum_{k=0}^{\infty}H_{I,I}^k\cdot R_ID_{\sqrt{\pi}}^{-1}]\|\cdot\|\mathcal{E}\| \\ &\leq \|\mathbb{E}_{\mathcal{G}}[D_{\sqrt{\pi}}^{-1}R_I^*И_{I,I}R_ID_{\sqrt{\pi}}^{-1}]\| + \|\mathbb{E}_{\mathcal{G}}[D_{\sqrt{\pi}}^{-1}R_I^*H_{I,I}R_ID_{\sqrt{\pi}}^{-1}]\|\cdot\frac{\varepsilon^2}{2} \quad (4.70) \\ &\qquad + \|\mathbb{E}[D_{\sqrt{\pi}}^{-1}R_I^*H_{I,I}H_{I,I}^*R_ID_{\sqrt{\pi}}^{-1}]\|^{1/2}\max_{I\in\mathcal{G}}\|\sum_{k=0}^{\infty}H_{I,I}^k\|\cdot \\ &\qquad\qquad \left(\|\mathbb{E}[D_{\sqrt{\pi}}^{-1}R_I^*И_{I,I}И_{I,I}^*R_ID_{\sqrt{\pi}}^{-1}]\|^{1/2} + \|\mathbb{E}[D_{\sqrt{\pi}}^{-1}R_I^*H_{I,I}H_{I,I}^*R_ID_{\sqrt{\pi}}^{-1}]\|^{1/2}\cdot\frac{\varepsilon^2}{2}\right) \\ &\overset{\star}{\leq} 6\|\Psi D_{\sqrt{\pi}}\|\|ZD_{\sqrt{\pi}}\| + 3\|\Psi D_{\sqrt{\pi}}\|^2\cdot\frac{\varepsilon^2}{2} \\ &\qquad + \sqrt{2}\|\Psi D_{\sqrt{\pi}}\|\cdot\frac{1}{(1-\vartheta)}\cdot\left(\frac{6}{8\sqrt{2}}\|ZD_{\sqrt{\pi}}\| + \frac{3}{2}\varepsilon\|\Psi D_{\sqrt{\pi}}\| + \sqrt{2}\|\Psi D_{\sqrt{\pi}}\|\frac{\varepsilon^2}{2}\right) \\ &\leq 7\|\Psi D_{\sqrt{\pi}}\|\|ZD_{\sqrt{\pi}}\| + 7\|\Psi D_{\sqrt{\pi}}\|^2\varepsilon, \quad (4.71)\end{aligned}$$

where in $(\star)$ we used 4.20(b), (4.67) and (4.68). Collecting all of the above and putting it back into (4.63) and (4.62) yields

$$\begin{aligned}\|\mathbb{E}[\hat{Y}]\| &\leq \mathbb{P}(\mathcal{H}^c)2\rho\underline{\alpha} + \mathbb{P}(\mathcal{G}^c)\rho\underline{\alpha} + \|D_\alpha^{-1}\| \cdot \|\Phi D_{\sqrt{\pi}}\| \cdot \left[7\|\Psi D_{\sqrt{\pi}}\|\|ZD_{\sqrt{\pi}}\| + 7\|\Psi D_{\sqrt{\pi}}\|^2\varepsilon\right] \\ &\leq \frac{1}{32}\delta_\star\underline{\alpha} + 14\|D_\alpha^{-1}\| \cdot \|\Phi D_{\sqrt{\pi}}\| \cdot \|\Psi D_{\sqrt{\pi}}\| \cdot \delta, \end{aligned} \tag{4.72}$$

where the bound on the probabilities follows from Lemma 4.11. Thus by the assumptions of our Proposition we have

$$\|\mathbb{E}[\hat{Y}]\| \leq \frac{1}{32}\delta_\star\underline{\alpha} + m_4\frac{\underline{\alpha}\gamma^2}{C^2\log(K\rho/\delta_\star)} \cdot \delta. \tag{4.73}$$

Also due to our assumption we always have $\delta \leq \sqrt{2}$ and so for $C \geq \sqrt{2} \cdot 256m_4$ we get

$$\|\mathbb{E}[\hat{Y}]\| \leq \frac{1}{32}\delta_\star\underline{\alpha} + \frac{1}{2}\min\left\{\frac{\underline{\alpha}\gamma^2}{128C\log(K\rho/\delta_\star)}, \frac{\underline{\alpha}\gamma^2}{128C\log(K\rho/\delta_\star)} \cdot \delta\right\} \leq \underline{\alpha}\Delta/2 =: m. \tag{4.74}$$

Finally an application of the matrix Bernstein inequality for $t = \underline{\alpha}\Delta/2$ with $R = \frac{3}{4}\underline{\alpha}\rho$ and $m = \underline{\alpha}\Delta/2$, and some simplifications yield the desired bound. ■

The next lemma is going to show that the matrix $B = \sum_{n=1}^N \hat{x}_n\hat{x}_n^*$ essentially behaves like a diagonal matrix. Together with the appropriate diagonal scaling matrices $D_{\sqrt{\pi}}, D_\beta$ and D_α we are going to show that

$$(D_{\sqrt{\pi}\cdot\alpha})^{-1}B(D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1} \approx \mathbb{I}. \tag{4.75}$$

For that we are again going to invoke the matrix Bernstein inequality. And again the main difficulty lies in calculating the expected value of the involved quantities.

Lemma 4.13 (proof of Claim 2) *Under the assumptions of Proposition 4.3 we have*

$$\mathbb{P}\left(\|(D_{\sqrt{\pi}\cdot\alpha})^{-1}B(D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1} - \mathbb{I}\| > \Delta\right) \leq (d+K)\exp\left(-\frac{N(\Delta/2)^2}{2\rho^2 + \rho\Delta/2}\right) \tag{4.76}$$

where

$$\rho = 2\kappa^2\|\Phi\|^2 S\gamma^{-2}\underline{\alpha}^{-2}\underline{\pi}^{-3/2}.$$

Proof To show the above statement, we are going to follow the approach in the result above very closely. The idea is to again write $(D_{\sqrt{\pi}\cdot\alpha})^{-1}B(D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1} - \mathbb{I}$ as a scaled sum of independent random matrices $\hat{Y}_n$ and apply matrix Bernstein to show that we indeed have concentration. Recalling that $\hat{I}_n$ denotes the set found by the thresholding and that $\mathcal{B}(v) := \{I : \|\Psi_I^\dagger v\| \leq \kappa\|v\|\}$ denotes the set of possible stable supports for v, we define for $n \in [N]$ the matrices $\hat{Y}_n$ as well as their auxiliary counterparts Y_n as

$$\begin{aligned} &\hat{Y}_n := (D_{\sqrt{\pi}\cdot\alpha})^{-1}R_{\hat{I}_n}^*\Psi_{\hat{I}_n}^\dagger y_n y_n^*\Psi_{\hat{I}_n}^{\dagger *}R_{\hat{I}_n}(D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1}\mathbb{1}_{\mathcal{B}(y_n)}(\hat{I}_n) - \mathbb{I} \\ \text{and} \quad &Y_n := (D_{\sqrt{\pi}\cdot\alpha})^{-1}R_{I_n}^*\Psi_{I_n}^\dagger y_n y_n^*\Psi_{I_n}^{\dagger *}R_{I_n}(D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1}\mathbb{1}_{\mathcal{B}(y_n)}(I_n) - \operatorname{diag}(\mathbf{1}_{I_n})D_\pi^{-1}. \end{aligned}$$

Both matrices can be bounded as

$$\max\{\|\hat{Y}_n\|, \|Y_n\|\} \leq \kappa^2\|y\|^2\|D_\alpha^{-2}\|\|D_\beta^{-1}\|\|D_\pi^{-1}\| + \|D_\pi^{-1}\| \leq \frac{3}{4}\rho =: R. \tag{4.77}$$

Further on $\mathcal{H}$, meaning whenever thresholding succeeds, the first terms of $\hat{Y}_n$ and Y_n again coincide while the second terms are the same in expectation, that is $\mathbb{E}[\mathrm{diag}(\mathbf{1}_{I_n})D_\pi^{-1}] = \mathbb{I}$. So with the same argument as in (4.62) and as usual dropping the index n for convenience, we get

$$\|\mathbb{E}[\hat{Y}]\| \leq 2\rho \cdot \mathbb{P}(\mathcal{H}^c) + \rho \cdot \mathbb{P}(\mathcal{G}^c) + \|\mathbb{E}[\mathbb{1}_{\mathcal{G}}(I)Y]\|. \tag{4.78}$$

Similarly as in (4.63) we next use that all well conditioned supports are stable for any signal y, meaning $\mathcal{G} \subseteq \mathcal{B}(y)$. Taking the expectation over (σ, c) therefore yields

$$\begin{aligned}
\|\mathbb{E}[\mathbb{1}_{\mathcal{G}}(I)Y]\| &= \|\mathbb{E}_I[\mathbb{1}_{\mathcal{G}}(I)\,\mathbb{E}_{\sigma,c}[(D_{\sqrt{\pi}\cdot\alpha})^{-1}R_I^*\Psi_I^\dagger\Phi_I x_I x_I^*\Phi_I^*\Psi_I^{\dagger *}R_I(D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1} - \mathrm{diag}(\mathbf{1}_I)D_\pi^{-1}]]\| \\
&= \|\mathbb{E}_{\mathcal{G}}[(D_{\sqrt{\pi}\cdot\alpha})^{-1}R_I^*\Psi_I^\dagger\Phi_I\Phi_I^*\Psi_I^{\dagger *}R_I(D_{\sqrt{\pi}\cdot\alpha})^{-1} - (D_{\sqrt{\pi}\cdot\alpha})^{-1}R_I^*R_I D_\alpha^2(D_{\sqrt{\pi}\cdot\alpha})^{-1}]\| \\
&\leq \|D_\alpha^{-2}\| \cdot \left\|\mathbb{E}_{\mathcal{G}}\left[D_{\sqrt{\pi}}^{-1}R_I^*\left(\Psi_I^\dagger\Phi_I\Phi_I^*\Psi_I^{\dagger *} - (D_\alpha)_{I,I}^2\right)R_I D_{\sqrt{\pi}}^{-1}\right]\right\|.
\end{aligned} \tag{4.79}$$

We next recall the shorthands $\mathit{И} := Z^*\Psi - \mathcal{E}$ and $H = \Psi^*\Psi - \mathbb{I}$, the identity $D_\alpha = \mathbb{I} - \mathcal{E}$ and that on $\mathcal{G}$ we have $\Psi_I^\dagger\Psi_I = \mathbb{I}_S$, which allows us to rewrite the expression inside the expectation above using again a Neumann series, (4.64), as

$$\begin{aligned}
\Psi_I^\dagger\Phi_I\Phi_I^*(\Psi_I^\dagger)^* - (D_\alpha^2)_{I,I} &= \Psi_I^\dagger(\Psi_I - Z_I)(\Psi_I^* - Z_I^*)(\Psi_I^\dagger)^* - (D_\alpha)_{I,I}^2 \\
&= \mathbb{I}_S - \Psi_I^\dagger Z_I - Z_I^*\Psi_I^{\dagger *} + \Psi_I^\dagger Z_I Z_I^*\Psi_I^{\dagger *} - (D_\alpha)_{I,I}^2 \\
&= \underbrace{\mathcal{E}_{I,I} - \textstyle\sum_{k=0}^\infty H_{I,I}^k\Psi_I^*Z_I}_{\text{Ia}} + \underbrace{\mathcal{E}_{I,I} - Z_I^*\Psi_I\textstyle\sum_{k=0}^\infty H_{I,I}^k}_{\text{Ib}} \\
&\qquad + \underbrace{\textstyle\sum_{k=0}^\infty H_{I,I}^k\Psi_I^*Z_IZ_I^*\Psi_I\sum_{k=0}^\infty H_{I,I}^k - \mathcal{E}_{I,I}^2}_{\text{II}}.
\end{aligned} \tag{4.80}$$

Due to the triangle inequality it suffices to estimate for each of the three terms the corresponding operator norm of its expectation as in (4.79). Since Term Ia is the transpose of Term Ib, it has the same corresponding operator norm, already bounded in (4.71) as

$$\|\mathbb{E}_{\mathcal{G}}[D_{\sqrt{\pi}}^{-1}R_I^*\Big(Z_I^*\Psi_I\sum_{k=0}^\infty H_{I,I}^k - \mathcal{E}_{I,I}\Big)R_I D_{\sqrt{\pi}}^{-1}]\| \leq 14\|D_\alpha^{-1}\|\|\Phi D_{\sqrt{\pi}}\|\|\Psi D_{\sqrt{\pi}}\| \cdot \delta. \tag{4.81}$$

Term II has to be further decomposed into

$$\begin{aligned}
\mathrm{II} = \underbrace{\Psi_I^*Z_IZ_I^*\Psi_I - \mathcal{E}_{I,I}^2}_{\text{IIa}} + \underbrace{\sum_{k\geq 1} H_{I,I}^k\Psi_I^*Z_IZ_I^*\Psi_I}_{\text{IIb}} + \underbrace{\Psi_I^*Z_IZ_I^*\Psi_I\sum_{k\geq 1}H_{I,I}^k}_{\text{IIc}} \\
+ \underbrace{\sum_{k\geq 1}H_{I,I}^k\Psi_I^*Z_IZ_I^*\Psi_I\sum_{k\geq 1}H_{I,I}^k}_{\text{IId}}.
\end{aligned}$$

The norm term corresponding to IIa can be straightforwardly bounded using (4.69), Corollary 4.20(b) and $\varepsilon \leq \sqrt{2}$ we get

$$\begin{aligned}
&\|\mathbb{E}_{\mathcal{G}}[D_{\sqrt{\pi}}^{-1}R_I^*(\Psi_I^*Z_IZ_I^*\Psi_I - \mathcal{E}_{I,I}^2)R_I D_{\sqrt{\pi}}^{-1}]\| \\
&\qquad \leq \|\mathbb{E}_{\mathcal{G}}[D_{\sqrt{\pi}}^{-1}R_I^*\mathit{И}_{I,I}^*\mathit{И}_{I,I}R_I D_{\sqrt{\pi}}^{-1}]\| + 2\|\mathbb{E}_{\mathcal{G}}[D_{\sqrt{\pi}}^{-1}R_I^*\mathit{И}_{I,I}^*\mathcal{E}_{I,I}R_I D_{\sqrt{\pi}}^{-1}]\| \\
&\qquad \leq 2\|ZD_{\sqrt{\pi}}\|^2 + 2\|\mathbb{E}_{\mathcal{G}}[D_{\sqrt{\pi}}^{-1}R_I^*\mathit{И}_{I,I}^*R_I D_{\sqrt{\pi}}^{-1}]\| \cdot \frac{\varepsilon^2}{2} \\
&\qquad \leq 2\|ZD_{\sqrt{\pi}}\|^2 + 6\|D_{\sqrt{\pi}}\mathit{И}^*D_{\sqrt{\pi}}\| \leq 2\|ZD_{\sqrt{\pi}}\|^2 + 12\|\Psi D_{\sqrt{\pi}}\|\|ZD_{\sqrt{\pi}}\|.
\end{aligned} \tag{4.82}$$

The terms IIb and IIc are again each other's transpose, so we only need to bound the corresponding norm for IIb. In order to do so we split

$$\Psi_I^* Z_I Z_I^* \Psi_I = \Psi_I^* Z_I \mathit{И}_{I,I} + \mathit{И}_{I,I}^* \mathcal{E}_{I,I} + \mathcal{E}_{I,I}^2.$$

Theorem 4.7 together with the bounds in (4.67-4.69) and the fact that on $\mathcal{G}$ we have $\|Z_I\| \leq \|\Phi_I\| + \|\Psi_I\| \leq 2\sqrt{1+\vartheta}$ then yields

$$\begin{aligned}
&\|\mathbb{E}_{\mathcal{G}}[D_{\sqrt{\pi}}^{-1} R_I^* H_{I,I} \sum_{k=0}^{\infty} H_{I,I}^k \Psi_I^* Z_I \mathit{И}_{I,I} R_I D_{\sqrt{\pi}}^{-1}\| \\
&\quad \leq \|\mathbb{E}[D_{\sqrt{\pi}}^{-1} R_I H_{I,I} H_{I,I}^* R_I^* D_{\sqrt{\pi}}^{-1}]\|^{\frac{1}{2}} \cdot \max_{I\in\mathcal{G}} \|\sum_{k=0}^{\infty} H_{I,I}^k \Psi_I^* Z_I\| \cdot \|\mathbb{E}[D_{\sqrt{\pi}}^{-1} R_I^* \mathit{И}_{I,I}^* \mathit{И}_{I,I} R_I D_{\sqrt{\pi}}^{-1}\|^{\frac{1}{2}} \\
&\quad \leq \sqrt{2}\|\Psi D_{\sqrt{\pi}}\| \cdot 2\frac{1+\vartheta}{1-\vartheta} \cdot \sqrt{2}\|Z D_{\sqrt{\pi}}\|, \qquad (4.83)
\end{aligned}$$

as well as

$$\begin{aligned}
&\|\mathbb{E}_{\mathcal{G}}[D_{\sqrt{\pi}}^{-1} R_I^* H_{I,I} \sum_{k=0}^{\infty} H_{I,I}^k \mathit{И}_{I,I}^* \mathcal{E}_{I,I} R_I D_{\sqrt{\pi}}^{-1}\| = \|\mathbb{E}_{\mathcal{G}}[D_{\sqrt{\pi}}^{-1} R_I^* H_{I,I} \sum_{k=0}^{\infty} H_{I,I}^k \mathit{И}_{I,I}^* R_I D_{\sqrt{\pi}}^{-1} \mathcal{E}\| \\
&\quad \leq \|\mathbb{E}[D_{\sqrt{\pi}}^{-1} R_I H_{I,I} H_{I,I}^* R_I^* D_{\sqrt{\pi}}^{-1}]\|^{\frac{1}{2}} \cdot \max_{I\in\mathcal{G}} \|\sum_{k=0}^{\infty} H_{I,I}^k\| \cdot \|\mathbb{E}[D_{\sqrt{\pi}}^{-1} R_I^* \mathit{И}_{I,I}^* \mathit{И}_{I,I} R_I D_{\sqrt{\pi}}^{-1}\|^{\frac{1}{2}} \cdot \|\mathcal{E}\| \\
&\quad \leq \sqrt{2}\|\Psi D_{\sqrt{\pi}}\| \cdot \frac{1}{1-\vartheta} \cdot \sqrt{2}\|Z D_{\sqrt{\pi}}\| \cdot \varepsilon^2/2, \qquad (4.84)
\end{aligned}$$

and finally using also Corollary 4.20(b)

$$\begin{aligned}
&\|\mathbb{E}_{\mathcal{G}}[D_{\sqrt{\pi}}^{-1} R_I^* H_{I,I} \sum_{k=0}^{\infty} H_{I,I}^k \mathcal{E}_{I,I}^2 R_I D_{\sqrt{\pi}}^{-1}\| \\
&\quad \leq \|\mathbb{E}_{\mathcal{G}}[D_{\sqrt{\pi}}^{-1} R_I^* H_{I,I} R_I D_{\sqrt{\pi}}^{-1}\| \cdot \|\mathcal{E}\|^2 + \|\mathbb{E}_{\mathcal{G}}[D_{\sqrt{\pi}}^{-1} R_I^* H_{I,I} \sum_{k=0}^{\infty} H_{I,I}^k H_{I,I} R_I D_{\sqrt{\pi}}^{-1}\| \cdot \|\mathcal{E}\|^2 \\
&\quad \leq 3\|\Psi D_{\sqrt{\pi}}\|^2 \cdot \|\mathcal{E}\|^2 + \|\mathbb{E}[D_{\sqrt{\pi}}^{-1} R_I H_{I,I} H_{I,I}^* R_I^* D_{\sqrt{\pi}}^{-1}]\| \cdot \max_{I\in\mathcal{G}} \|\sum_{k=0}^{\infty} H_{I,I}^k\| \cdot \|\mathcal{E}\|^2 \\
&\quad \leq 3\|\Psi D_{\sqrt{\pi}}\|^2 \cdot \varepsilon^4/4 + 2\|\Psi D_{\sqrt{\pi}}\|^2 \cdot (1-\vartheta)^{-1} \cdot \varepsilon^4/4. \qquad (4.85)
\end{aligned}$$

Combining (4.83-4.85) and assuming that $\vartheta \leq 1/4$ yields the following bound for the norm term corresponding to IIb

$$\begin{aligned}
&\|\mathbb{E}_{\mathcal{G}}[D_{\sqrt{\pi}}^{-1} R_I^* \sum_{k=1}^{\infty} H_{I,I}^k \Psi_I^* Z_I Z_I^* \Psi_I R_I D_{\sqrt{\pi}}^{-1}]\| \\
&\qquad \leq \frac{20}{3}\|\Psi D_{\sqrt{\pi}}\|\|Z D_{\sqrt{\pi}}\| + \frac{8}{3}\|\Psi D_{\sqrt{\pi}}\|\|Z D_{\sqrt{\pi}}\| \cdot \frac{\varepsilon^2}{2} + \frac{17}{3}\|\Psi D_{\sqrt{\pi}}\|^2 \cdot \frac{\varepsilon^4}{4} \\
&\qquad \leq 10\|\Psi D_{\sqrt{\pi}}\|\|Z D_{\sqrt{\pi}}\| + 5\|\Psi D_{\sqrt{\pi}}\|^2 \varepsilon. \qquad (4.86)
\end{aligned}$$

Before we bound the norm term corresponding to IId, note that on $\mathcal{G}$ we have two possibilities to bound $\|Z_I Z_I^*\| = \|Z_I\|^2$, either via $\|Z_I\| \leq \|\Psi_I\| + \|\Phi_I\| \leq 2\sqrt{1+\vartheta}$ or simply using that by definition $\|Z_I\|$ is close to δ on $\mathcal{G}$, cp. (4.47). Combining both estimates we get

$$\max_{I\in\mathcal{G}} \|Z_I Z_I^*\| \leq 2\sqrt{1+\vartheta} \cdot \min\left\{2\sqrt{1+\vartheta}, \sqrt{2e^2 \log(320K\rho/\delta_\star)} \cdot \delta\right\}. \qquad (4.87)$$

Applying this together with Theorem 4.7 yields

$$\begin{aligned}
&\|\mathbb{E}_{\mathcal{G}}[D_{\sqrt{\pi}}^{-1}R_I^*\sum_{k=1}^{\infty}H_{I,I}^k\Psi_I^*Z_IZ_I^*\Psi_I\sum_{k=1}^{\infty}H_{I,I}^kR_ID_{\sqrt{\pi}}^{-1}]\| \\
&\qquad\le \|\mathbb{E}[D_{\sqrt{\pi}}^{-1}R_IH_{I,I}H_{I,I}^*R_I^*D_{\sqrt{\pi}}^{-1}]\|\cdot\max_{I\in\mathcal{G}}\|\sum_{k=0}^{\infty}H_{I,I}^k\Psi_I^*Z_IZ_I^*\Psi_I\sum_{k=0}^{\infty}H_{I,I}^k\| \\
&\qquad\le 2\|\Psi D_{\sqrt{\pi}}\|^2\cdot\frac{1+\vartheta}{(1-\vartheta)^2}\cdot 2\sqrt{1+\vartheta}\min\left\{2\sqrt{1+\vartheta},\sqrt{2e^2\log(320K\rho/\delta_\star)}\cdot\delta\right\} \\
&\qquad\le 10\|\Psi D_{\sqrt{\pi}}\|^2\cdot\min\left\{3,\sqrt{2e^2\log(320K\rho/\delta_\star)}\cdot\delta\right\}. \qquad (4.88)
\end{aligned}$$

Combining the bounds for the norm terms corresponding to Ia/b and IIa-d collected from (4.81), (4.82),(4.86) and (4.88) into a bound for $\|\mathbb{E}[\mathbb{1}_{\mathcal{G}}(I)Y]\|$ in (4.79), which in turn is substituted into (4.78) with the same bound for $\mathbb{P}(\mathcal{H}^c)\,2\rho+\mathbb{P}(\mathcal{G}^c)\,\rho\le\frac{1}{32}\delta_\star$ as in the last Lemma finally yields

$$\begin{aligned}
\|\mathbb{E}[\hat{Y}]\|\le\frac{1}{32}\delta_\star+\|D_\alpha^{-2}\|\Big[&2\|ZD_{\sqrt{\pi}}\|^2+28\|\Phi D_{\sqrt{\pi}}\|\|\Psi D_{\sqrt{\pi}}\|\cdot\delta+32\|\Psi D_{\sqrt{\pi}}\|\|ZD_{\sqrt{\pi}}\| \\
&+10\|\Psi D_{\sqrt{\pi}}\|^2\varepsilon+10\|\Psi D_{\sqrt{\pi}}\|^2\cdot\min\left\{3,\sqrt{2e^2\log(320K\rho/\delta_\star)}\cdot\delta\right\}\Big]. \qquad (4.89)
\end{aligned}$$

We can use either $\|ZD_{\sqrt{\pi}}\|\le\|\Psi D_{\sqrt{\pi}}\|+\|\Phi D_{\sqrt{\pi}}\|$ or $\|ZD_{\sqrt{\pi}}\|+\|\Psi D_{\sqrt{\pi}}\|\varepsilon\lesssim\delta$ together with $\|\Psi D_{\sqrt{\pi}}\|\le\frac{1}{C\sqrt{\log(K\rho/\delta_\star)}}$ to get

$$\begin{aligned}
\|\mathbb{E}[\hat{Y}]\|&\le\frac{1}{32}\delta_\star+m_5\|D_\alpha^{-2}\|\min\left\{\max\{\|\Psi D_{\sqrt{\pi}}\|^2,\|\Phi D_{\sqrt{\pi}}\|^2\},\max\{\|\Psi D_{\sqrt{\pi}}\|,\|\Phi D_{\sqrt{\pi}}\|\}\cdot\delta\right\} \\
&\le\frac{1}{32}\delta_\star+m_6\min\left\{\frac{\gamma^2}{C^2\log(K\rho/\delta_\star)},\frac{\gamma}{C\underline{\alpha}\sqrt{\log(K\rho/\delta_\star)}}\cdot\delta\right\}. \qquad (4.90)
\end{aligned}$$

Recall that $\underline{\alpha}=1-\varepsilon^2/2\le 1$. The second term of this minimum will be attained only if $\delta\lesssim\frac{1}{C\sqrt{\log(K\rho/\delta_\star)}}$ in which case we surely have $\underline{\alpha}\ge 1/2$. Thus the inequality gets weaker if we multiply the second term by 2α to get

$$\|\mathbb{E}[\hat{Y}]\|\le\frac{1}{32}\delta_\star+m_7\min\left\{\frac{\gamma^2}{C^2\log(K\rho/\delta_\star)},\frac{\gamma}{C\sqrt{\log(K\rho/\delta_\star)}}\cdot\delta\right\}. \qquad (4.91)$$

So for $C\ge 256m_7$,

$$\|\mathbb{E}[\hat{Y}]\|\le\frac{1}{32}\delta_\star+\frac{1}{2}\min\left\{\frac{\gamma^2}{128C\log(K\rho/\delta_\star)},\frac{\gamma}{128\sqrt{\log(K\rho/\delta_\star)}}\cdot\delta\right\}=\frac{1}{2}\Delta=:m. \qquad (4.92)$$

As before an application of the matrix Bernstein inequality for $t=\Delta/2$ with $R=\frac{3}{4}\rho$ and $m=\Delta/2$, and some simplifications yield the desired bound. ∎

Now we turn to bounding individual columns of the random matrices we have to deal with. This will again be done by applying a Bernstein-type inequality. The main difficulty again lies in calculating the expected value of the involved terms.

Lemma 4.14 (proof of Claim 3) *Assume the conditions of Proposition 4.3, then*

$$\mathbb{P}\Big(\|\Phi A(D_{\pi\cdot\alpha\cdot\beta})^{-1}e_\ell-\phi_\ell\|>\underline{\alpha}\Delta\Big)\le 28\exp\left(-\frac{N(\Delta/2)^2}{2\rho^2+\rho\Delta/2}\right)$$

where

$$\rho = 2\kappa^2\|\Phi\|^2 S\gamma^{-2}\underline{\alpha}^{-2}\underline{\pi}^{-3/2}.$$

Proof As in the matrix case before the idea is to write the vector whose norm we want to estimate as sum of independent random vectors based on the signals y_n and use Bernstein inequality. To this end we define for a fixed index ℓ the random vectors

$$\begin{aligned}\hat{Y}_n &:= \Big(y_n y_n^* \Phi_{I_n}^* \Psi_{\hat{I}_n}^\dagger R_{\hat{I}_n} (D_{\pi\cdot\alpha\cdot\beta})^{-1} \mathbb{1}_{\mathcal{B}(y_n)}(\hat{I}_n) - \Phi\Big) e_\ell,\\ Y_n &:= \Big(y_n y_n^* (\Psi_{I_n}^\dagger)^* R_{I_n} (D_{\pi\cdot\alpha\cdot\beta})^{-1} \cdot \mathbb{1}_{\mathcal{B}(y_n)}(I_n) - \Phi \operatorname{diag}(\mathbf{1}_{I_n}) D_\pi^{-1}\Big) e_\ell\end{aligned}$$

Note that we can obtain $\hat{Y}_n, Y_n$ by multiplying the analogue matrices in the proof of Lemma 4.12 (Claim 1) from the left by $D_{\sqrt{\pi}}^{-1} e_\ell$. Following the proof strategy of Lemma 4.12 with the necessary changes, we first bound the ℓ_2-norm of the random vectors $\hat{Y}_n, Y_n$ as

$$\max\{\|\hat{Y}_n\|, \|Y_n\|\} \le \kappa\|\Phi\|^2 S c_{\max}^2 \|\|D_\pi^{-1}\|\|D_\beta^{-1}\|\|D_\alpha^{-1}\| + 1 \le \frac{3}{4}\underline{\alpha}\rho =: R,$$

while, repeating the procedures in (4.62) and (4.63), we get for the expectation,

$$\begin{aligned}\|\mathbb{E}[\hat{Y}]\| &\le \mathbb{P}(\mathcal{H}^c) 2\rho\underline{\alpha} + \mathbb{P}(\mathcal{G}^c)\rho\underline{\alpha} + \|\mathbb{E}[\mathbb{1}_{\mathcal{G}}(I) Y]\|\\ &\le \mathbb{P}(\mathcal{H}^c) 2\rho\underline{\alpha} + \mathbb{P}(\mathcal{G}^c)\rho\underline{\alpha} + \|D_\alpha^{-1}\| \cdot \big\|\Phi\mathbb{E}_{\mathcal{G}}\big[R_I^*\big(\Phi_I^*\Psi_I^{\dagger *} - (D_\alpha)_{I,I}\big) R_I D_\pi^{-1}\big] e_\ell\big\|\\ &= \mathbb{P}(\mathcal{H}^c) 2\rho\underline{\alpha} + \mathbb{P}(\mathcal{G}^c)\rho\underline{\alpha} + \|D_\alpha^{-1}\| \cdot \pi_\ell^{-1} \cdot \big\|\Phi\mathbb{E}_{\mathcal{G}}\big[R_I^*\big(\Phi_I^*\Psi_I^{\dagger *} - (D_\alpha)_{I,I}\big) R_I e_\ell\big]\big\|. \qquad (4.93)\end{aligned}$$

Using the decomposition $\mathbb{I} = \mathbb{I}_{\ell^c} + e_\ell e_\ell^*$ we split the norm into two parts

$$\begin{aligned}&\big\|\Phi\mathbb{E}_{\mathcal{G}}\big[R_I^*\big(\Phi_I^*\Psi_I^{\dagger *} - (D_\alpha)_{I,I}\big) R_I e_\ell\big]\big\|\\ &\quad\le \big\|\Phi\cdot\mathbb{I}_{\ell^c}\cdot\mathbb{E}_{\mathcal{G}}\big[R_I^*\big(\Phi_I^*\Psi_I^{\dagger *} - (D_\alpha)_{I,I}\big) R_I e_\ell\big]\big\|\\ &\qquad\qquad + \big\|\Phi\cdot e_\ell e_\ell^*\cdot\mathbb{E}_{\mathcal{G}}\big[R_I^*\big(\Phi_I^*\Psi_I^{\dagger *} - (D_\alpha)_{I,I}\big) R_I e_\ell\big]\big\|\\ &\quad\le \|\Phi D_{\sqrt{\pi}}\|\big\|\mathbb{E}_{\mathcal{G}}\big[D_{\sqrt{\pi}}^{-1}\mathbb{I}_{\ell^c} R_I^*\big(\Phi_I^*\Psi_I^{\dagger *} - (D_\alpha)_{I,I}\big) R_I e_\ell\big]\big\|\\ &\qquad\qquad + \big\|\mathbb{E}_{\mathcal{G}}\big[e_\ell^* R_I^*\big(\Phi_I^*\Psi_I^{\dagger *} - (D_\alpha)_{I,I}\big) R_I e_\ell\big]\big\|. \qquad (4.94)\end{aligned}$$

By the same decomposition as in (4.66) we have

$$\Phi_I^*(\Psi_I^\dagger)^* - (D_\alpha)_{I,I} = Z_I^*\Psi_I \sum_{k\ge 0} H_{I,I}^k - \mathcal{E}_{I,I}.$$

Note that for any diagonal matrix D we have $\mathbb{I}_{\ell^c} D e_\ell = 0$. So using the decomposition above, Theorem 4.7 and Corollary 4.20(d/k/f) together with $\vartheta \le \frac{1}{4}$, and $\pi_\ell \le \frac{1}{3}$ we can bound the first expectation in (4.94) as

$$\begin{aligned}&\|\mathbb{E}_{\mathcal{G}}[D_{\sqrt{\pi}}^{-1}\mathbb{I}_{\ell^c} R_I^*(Z_I^*\Psi_I \sum_{k\ge 0} H_{I,I}^k - \mathcal{E}_{I,I}) R_I e_\ell]\| = \|\mathbb{E}_{\mathcal{G}}[D_{\sqrt{\pi}}^{-1}\mathbb{I}_{\ell^c} R_I^* Z_I^*\Psi_I \sum_{k\ge 0} H_{I,I}^k R_I e_\ell]\|\\ &\quad\le \|\mathbb{E}_{\mathcal{G}}[D_{\sqrt{\pi}}^{-1}\mathbb{I}_{\ell^c} R_I^* Z_I^* \mathbb{1}_I(\ell)]\,\psi_\ell\| + \|\mathbb{E}_{\mathcal{G}}[D_{\sqrt{\pi}}^{-1}\mathbb{I}_{\ell^c} R_I^* Z_I^*\Psi_I \cdot \sum_{k\ge 0} H_{I,I}^k \cdot H_{I,\ell}\cdot \mathbb{1}_I(\ell)]\|\\ &\quad\le \|\mathbb{E}_{\mathcal{G}}[D_{\sqrt{\pi}}^{-1}\mathbb{I}_{\ell^c} R_I^* Z_I^*\cdot\mathbb{1}_I(\ell)]\| + \|\mathbb{E}[D_{\sqrt{\pi}}^{-1}\mathbb{I}_{\ell^c} R_I^* Z_I^*\Psi_I\Psi_I^* Z_I R_I \mathbb{I}_{\ell^c} D_{\sqrt{\pi}}^{-1}\cdot\mathbb{1}_I(\ell)]\|^{\frac{1}{2}}\\ &\qquad\qquad \cdot \max_{I\in\mathcal{G}}\|\sum_{k\ge 0} H_{I,I}^k\| \cdot \|\mathbb{E}[(H_{I,\ell})^* H_{I,\ell}\cdot\mathbb{1}_I(\ell)]\|^{\frac{1}{2}}\\ &\quad\le \pi_\ell\cdot\|ZD_{\sqrt{\pi}}\| + \sqrt{12\pi_\ell}\cdot\max\{\varepsilon, \|ZD_{\sqrt{\pi}}\|\}\cdot(1-\vartheta)^{-1}\cdot\sqrt{\pi_\ell}\cdot\|\Psi D_{\sqrt{\pi}}\|\\ &\quad\le \pi_\ell\cdot\left(\|ZD_{\sqrt{\pi}}\| + 5\delta\|\Psi D_{\sqrt{\pi}}\|\right). \qquad (4.95)\end{aligned}$$

To bound the second expectation in (4.94), we proceed very similarly, this time using Corollary 4.20 (e/f/i)

$$\begin{aligned}
&\|\mathbb{E}_{\mathcal{G}}[e_\ell^* \cdot R_I^*(Z_I^*\Psi_I \sum_{k\geq 0} H_{I,I}^k - \mathcal{E}_{I,I})R_I \cdot e_\ell]\| \\
&\quad \leq \|\mathbb{E}_{\mathcal{G}}[\underbrace{e_\ell^* \cdot R_I^*(Z_I^*\Psi_I - \mathcal{E}_{I,I})R_I \cdot e_\ell}_{=0}]\| + \|\mathbb{E}_{\mathcal{G}}[z_\ell^* \cdot \Psi_I \sum_{k\geq 0} H_{I,I}^k H_{I,\ell} \cdot \mathbb{1}_I(\ell)]\| \\
&\quad \leq \|z_\ell\| \cdot \left(\|\mathbb{E}_{\mathcal{G}}[\Psi_I H_{I,\ell} \cdot \mathbb{1}_I(\ell)]\| + \|\mathbb{E}_{\mathcal{G}}[\Psi_I H_{I,I} \sum_{k\geq 0} H_{I,I}^k H_{I,\ell} \cdot \mathbb{1}_I(\ell)]\|\right) \\
&\quad \leq \varepsilon \cdot \pi_\ell \cdot \|\Psi D_{\sqrt{\pi}}\|^2 \\
&\qquad + \varepsilon \cdot \|\mathbb{E}[\Psi_I H_{I,I} H_{I,I}^* \Psi_I^* \cdot \mathbb{1}_I(\ell)]\|^{\frac{1}{2}} \cdot \max_{I\in\mathcal{G}} \|\sum_{k\geq 0} H_{I,I}^k\| \cdot \|\mathbb{E}[(H_{I,\ell})^* H_{I,\ell} \cdot \mathbb{1}_I(\ell)]\|^{\frac{1}{2}} \\
&\quad \leq \varepsilon \cdot \pi_\ell \cdot \|\Psi D_{\sqrt{\pi}}\|^2 + \varepsilon \cdot \sqrt{2\pi_\ell} \cdot \|\Psi D_{\sqrt{\pi}}\| \cdot (1-\vartheta)^{-1} \cdot \sqrt{\pi_\ell} \cdot \|\Psi D_{\sqrt{\pi}}\| \\
&\quad \leq 3\varepsilon \cdot \pi_\ell \cdot \|\Psi D_{\sqrt{\pi}}\|^2
\end{aligned}$$

Plugging these bounds back into (4.94) and (4.93) yields

$$\begin{aligned}
\|\mathbb{E}[\hat{Y}]\| \leq\ & \mathbb{P}\left(\mathcal{H}^c\right) 2\rho\underline{\alpha} + \mathbb{P}\left(\mathcal{G}^c\right)\rho\underline{\alpha} \\
&+ \|D_\alpha^{-1}\| \cdot \left(\|ZD_{\sqrt{\pi}}\|\|\Phi D_{\sqrt{\pi}}\| + 3\varepsilon \cdot \|\Psi D_{\sqrt{\pi}}\|^2 + 5 \cdot \|\Phi D_{\sqrt{\pi}}\|\|\Psi D_{\sqrt{\pi}}\| \cdot \delta\right).
\end{aligned}$$

Using Lemma 4.11 to bound the first two terms involving the probabilities and using either $\|ZD_{\sqrt{\pi}}\| \leq \|\Psi D_{\sqrt{\pi}}\| + \|\Phi D_{\sqrt{\pi}}\|$ or $\|ZD_{\sqrt{\pi}}\| \leq \delta$ yields together with the assumptions of the proposition

$$\begin{aligned}
\|\mathbb{E}[\hat{Y}]\| &\leq \frac{1}{32}\delta_\star\underline{\alpha} + m_8\|D_\alpha^{-1}\| \min\left\{ \max\left\{\|\Psi D_{\sqrt{\pi}}\|^2, \|\Phi D_{\sqrt{\pi}}\|^2\right\}, \max\left\{\|\Psi D_{\sqrt{\pi}}\|, \|\Phi D_{\sqrt{\pi}}\|\right\} \cdot \delta \right\} \\
&\leq \frac{1}{32}\delta_\star\underline{\alpha} + m_9 \min\left\{ \frac{\underline{\alpha}\gamma^2}{C^2\log(K\rho/\delta_\star)}, \frac{\gamma}{C\sqrt{\log(K\rho/\delta_\star)}} \cdot \delta \right\}.
\end{aligned}$$

Recall that $\underline{\alpha} = 1 - \varepsilon^2/2 \leq 1$. The second term of this minimum will be attained only if $\delta \lesssim \frac{1}{C\sqrt{\log(K\rho/\delta_\star}}$ in which case we surely have $\underline{\alpha} \geq 1/2$. thus the inequality gets weaker if we multiply the second term by 2α to get

$$\begin{aligned}
\|\mathbb{E}[\hat{Y}]\| &\leq \frac{1}{32}\delta_\star\underline{\alpha} + m_{10} \min\left\{ \frac{\underline{\alpha}\gamma^2}{C^2\log(K\rho/\delta_\star)}, \frac{\underline{\alpha}\gamma}{C\sqrt{\log(K\rho/\delta_\star)}} \cdot \delta \right\} \\
&\leq \frac{1}{32}\delta_\star\underline{\alpha} + \frac{1}{2} \min\left\{ \frac{\underline{\alpha}\gamma^2}{128C\log(K\rho/\delta_\star)}, \frac{\underline{\alpha}\gamma}{128\sqrt{\log(K\rho/\delta_\star)}} \cdot \delta \right\} = \underline{\alpha}\Delta/2 := m, \quad (4.96)
\end{aligned}$$

where we used that $C \leq 256m_{10}$. Finally an application of the vector Bernstein inequality for $t = \underline{\alpha}\Delta/2$ with $R = \frac{3}{4}\underline{\alpha}\rho$ and $m = \underline{\alpha}\Delta/2$, and some simplifications yield the desired bound. ■

Now to the grand final, showing that Claim $C4$ is satisfied with high probability.

Lemma 4.15 (proof of Claim 4) *Assume the conditions of Proposition 4.3 then for* $\Lambda := \frac{\underline{\alpha}\gamma}{C\sqrt{\log(K\rho/\delta_\star)}}$

$$\mathbb{P}\left(\Lambda \cdot \|\mathbb{I}_{\ell^c}\left(D_{\sqrt{\pi}\cdot\alpha}\right)^{-1} B (D_{\pi\cdot\alpha\cdot\beta})^{-1} e_\ell\| > \Delta\right) \leq 28\exp\left(-\frac{N(\Delta/2)^2}{2\rho^2 + \rho\Delta/2}\right), \quad (4.97)$$

where

$$\rho = 2\kappa^2\|\Phi\|^2 S\gamma^{-2}\underline{\alpha}^{-2}\underline{\pi}^{-3/2}.$$

Proof As usual we rewrite the vector to bound as sum of random vectors based on the signals y_n and use Bernstein's inequality. Thus we define

$$\hat{Y}_n := \Lambda \cdot \mathbb{I}_{\ell^c} (D_{\sqrt{\pi}\cdot\alpha})^{-1} R^*_{\hat{I}_n} \Psi^\dagger_{\hat{I}_n} y_n y_n^* \Psi^{\dagger *}_{\hat{I}_n} R_{\hat{I}_n} (D_{\pi\cdot\alpha\cdot\beta})^{-1} \mathbb{1}_{\mathcal{B}(y_n)}(\hat{I}_n)\, e_\ell$$

and its counterpart Y_n by simply replacing in the above $\hat{I}_n$ by I_n. Since for any diagonal matrix D we have $\mathbb{I}_{\ell^c} D\, e_\ell = 0$ we can again obtain $\hat{Y}_n, Y_n$ from the corresponding matrices in the proof of Lemma 4.13 (Claim 2), this time by multiplying from the right by $\mathbb{I}_{\ell^c}$ and from the left by $D^{-1}_{\sqrt{\pi}} e_\ell$. Following the usual proof strategy we first bound the ℓ_2-norm of the random vectors $\hat{Y}_n, Y_n$ as

$$\max\{\|\hat{Y}_n\|, \|Y_n\|\} \le \kappa^2 \|y\|^2 \|D_\alpha^{-2}\| \|D_\beta^{-1}\| \|D_\pi^{-3/2}\| \le \frac{3}{4}\rho =: R,$$

while for the expectation we get similar to (4.79) and (4.93)

$$\begin{aligned}\|\mathbb{E}[\hat{Y}]\| &\le \mathbb{P}(\mathcal{H}^c) 2\rho + \mathbb{P}(\mathcal{G}^c)\rho + \|\mathbb{E}[\mathbb{1}_{\mathcal{G}}(I) Y]\| \\ &\le \mathbb{P}(\mathcal{H}^c) 2\rho + \mathbb{P}(\mathcal{G}^c)\rho + \Lambda \cdot \pi_\ell^{-1} \cdot \|D_\alpha^{-2}\| \cdot \big\|\mathbb{E}_{\mathcal{G}}\big[D^{-1}_{\sqrt{\pi}} \mathbb{I}_{\ell^c} R_I^*\big(\Psi_I^\dagger \Phi_I \Phi_I^* \Psi_I^{\dagger *}\big) R_I e_\ell\big]\big\|. \quad (4.98)\end{aligned}$$

Using a decomposition as in (4.80) and recalling that $\mathbb{I}_{\ell^c} D\, e_\ell = 0$, we get

$$\begin{aligned}&\mathbb{I}_{\ell^c} R_I^*\big(\Psi_I^\dagger \Phi_I \Phi_I^* \Psi_I^{\dagger *}\big) R_I e_\ell \\ &\quad = \mathbb{I}_{\ell^c} R_I^* \Bigg(\underbrace{Z_I^* \Psi_I \sum_{k\ge 0} H_{I,I}^k}_{\text{Ia}} + \underbrace{\sum_{k\ge 0} H_{I,I}^k \Psi_I^* Z_I}_{\text{Ib}} + \underbrace{\sum_{k\ge 0} H_{I,I}^k \Psi_I^* Z_I Z_I^* \Psi_I \sum_{k\ge 0} H_{I,I}^k}_{\text{II}} \Bigg) R_I e_\ell. \quad (4.99)\end{aligned}$$

The expectation corresponding to Term Ia has already been bounded in (4.95) so we move on to Term Ib. Using Theorem 4.7 and Corollary 4.20 (d/h) we get

$$\begin{aligned}&\|\mathbb{E}_{\mathcal{G}}[D^{-1}_{\sqrt{\pi}}\, \mathbb{I}_{\ell^c}\, R_I^* \sum_{k\ge 0} H_{I,I}^k \Psi_I^* Z_I R_I\, e_\ell]\| \\ &\quad \le \|\mathbb{E}_{\mathcal{G}}[D^{-1}_{\sqrt{\pi}} \mathbb{I}_{\ell^c} R_I^* \Psi_I^* z_\ell \cdot \mathbb{1}_I(\ell)]\| + \|\mathbb{E}_{\mathcal{G}}[D^{-1}_{\sqrt{\pi}} \mathbb{I}_{\ell^c} R_I^* H_{I,I} \sum_{k\ge 0} H_{I,I}^k \Psi_I^* z_\ell \cdot \mathbb{1}_I(\ell)]\| \\ &\quad \le \|\mathbb{E}_{\mathcal{G}}[D^{-1}_{\sqrt{\pi}} \mathbb{I}_{\ell^c} R_I^* \Psi_I^* \cdot \mathbb{1}_I(\ell)]\| \cdot \|z_\ell\| \\ &\qquad + \|\mathbb{E}[D^{-1}_{\sqrt{\pi}} \mathbb{I}_{\ell^c} R_I^* H_{I,I} H_{I,I}^* R_I \mathbb{I}_{\ell^c} D^{-1}_{\sqrt{\pi}} \mathbb{1}_I(\ell)]\|^{\frac{1}{2}} \cdot \max_{I\in\mathcal{G}} \|\sum_{k\ge 0} H_{I,I}^k \Psi_I^*\| \cdot \|\mathbb{E}[z_\ell^* z_\ell \cdot \mathbb{1}_I(\ell)]\|^{\frac{1}{2}} \\ &\quad \le \pi_\ell \cdot \|\Psi D_{\sqrt{\pi}}\| \cdot \varepsilon + 3 \cdot \sqrt{\pi_\ell} \cdot \max\{\mu(\Psi), \|\Psi D_{\sqrt{\pi}}\|\} \cdot \frac{\sqrt{1+\vartheta}}{1-\vartheta} \cdot \sqrt{\pi_\ell} \cdot \varepsilon \\ &\quad \le 6 \cdot \pi_\ell \cdot \max\{\mu(\Psi), \|\Psi D_{\sqrt{\pi}}\|\} \cdot \varepsilon.\end{aligned}$$

As probably feared Term II requires further decomposition

$$\text{II} = \underbrace{\Psi_I^* Z_I Z_I^* \Psi_I}_{\text{IIa}} + \underbrace{\sum_{k\ge 1} H_{I,I}^k \Psi_I^* Z_I Z_I^* \Psi_I}_{\text{IIb}} + \underbrace{\Psi_I^* Z_I Z_I^* \Psi_I \sum_{k\ge 1} H_{I,I}^k}_{\text{IIc}} + \underbrace{\sum_{k\ge 1} H_{I,I}^k \Psi_I^* Z_I Z_I^* \Psi_I \sum_{k\ge 1} H_{I,I}^k}_{\text{IId}}.$$

Looking at the first of the four terms above, Term IIa, we get using $Z_I^* = R_I Z^* = R_I(e_\ell e_\ell^* + \mathbb{I}_{\ell^c})Z^*$ and Corollary 4.20 (d,k,g)

$$\begin{aligned}
&\|\mathbb{E}_{\mathcal{G}}[D_{\sqrt{\pi}}^{-1}\mathbb{I}_{\ell^c}R_I^*\Psi_I^*Z_IZ_I^*\Psi_IR_I\,e_\ell]\| \\
&\quad= \|\mathbb{E}_{\mathcal{G}}[D_{\sqrt{\pi}}^{-1}\mathbb{I}_{\ell^c}R_I^*\Psi_I^*Z_IR_I(e_\ell e_\ell^* + \mathbb{I}_{\ell^c})Z^*\psi_\ell\cdot\mathbb{1}_I(\ell)]\| \\
&\quad\le \|\mathbb{E}_{\mathcal{G}}[D_{\sqrt{\pi}}^{-1}\mathbb{I}_{\ell^c}R_I^*\Psi_I^*z_\ell z_\ell^*\psi_\ell\cdot\mathbb{1}_I(\ell)]\| + \|\mathbb{E}_{\mathcal{G}}[D_{\sqrt{\pi}}^{-1}R_I^*\Psi_I^*Z_IR_I\mathbb{I}_{\ell^c}Z^*\psi_\ell\cdot\mathbb{1}_I(\ell)]\| \\
&\quad\le \|\mathbb{E}_{\mathcal{G}}[D_{\sqrt{\pi}}^{-1}\mathbb{I}_{\ell^c}R_I^*\Psi_I^*\cdot\mathbb{1}_I(\ell)]\|\cdot\|z_\ell z_\ell^*\psi_\ell\| + \|\mathbb{E}_{\mathcal{G}}[D_{\sqrt{\pi}}^{-1}\mathbb{I}_{\ell^c}R_I^*\Psi_I^*Z_I\cdot R_I\mathbb{I}_{\ell^c}Z^*\cdot\mathbb{1}_I(\ell)]\|\cdot\|\psi_\ell\| \\
&\quad\le \pi_\ell\cdot\|\Psi D_{\sqrt{\pi}}\|\cdot\frac{\varepsilon^3}{2} + \|\mathbb{E}[D_{\sqrt{\pi}}^{-1}\mathbb{I}_{\ell^c}R_I^*\Psi_I^*Z_IZ_I^*\Psi_I\mathbb{I}_{\ell^c}D_{\sqrt{\pi}}^{-1}\mathbb{1}_I(\ell)]\|^{\frac12}\cdot\|\mathbb{E}[Z\mathbb{I}_{\ell^c}R_I^*R_I\mathbb{I}_{\ell^c}Z^*\mathbb{1}_I(\ell)]\|^{\frac12} \\
&\quad\le \pi_\ell\cdot\|\Psi D_{\sqrt{\pi}}\|\cdot\frac{\varepsilon^3}{2} + \sqrt{\pi_\ell}\cdot\max\{\varepsilon,\|ZD_{\sqrt{\pi}}\|\}\sqrt{12}\cdot\sqrt{\pi_\ell}\cdot\|ZD_{\sqrt{\pi}}\| \\
&\quad\le \pi_\ell\cdot\Big(\|\Psi D_{\sqrt{\pi}}\|\cdot\frac{\varepsilon^3}{2} + \sqrt{12}\cdot\delta\cdot\|ZD_{\sqrt{\pi}}\|\Big),
\end{aligned}$$

where we applied Corollary 4.20 (g) to $V = Z\mathbb{I}_{\ell^c}$, meaning $V_I = Z\mathbb{I}_{\ell^c}R_I^*$ and $v_\ell = 0$.
Term IIb will be treated the same way. We apply Theorem 4.7 and Corollary 4.20(h,g) together with $\vartheta \le \frac14$ and the fact that on $\mathcal{G}$ we have $\|Z_I\| \le \|\Psi_I\| + \|\Phi_I\| \le 2\sqrt{1+\vartheta}$ to get

$$\begin{aligned}
&\|\mathbb{E}_{\mathcal{G}}[D_{\sqrt{\pi}}^{-1}\mathbb{I}_{\ell^c}R_I^*\sum_{k\ge1}H_{I,I}^k\Psi_I^*Z_IZ_I^*\Psi_IR_I\,e_\ell]\| \\
&\qquad= \|\mathbb{E}_{\mathcal{G}}[D_{\sqrt{\pi}}^{-1}\mathbb{I}_{\ell^c}R_I^*H_{I,I}\cdot\sum_{k\ge0}H_{I,I}^k\Psi_I^*Z_I\cdot Z_I^*\psi_\ell\mathbb{1}_I(\ell)]\| \\
&\qquad\le \|\mathbb{E}_{\mathcal{G}}[D_{\sqrt{\pi}}^{-1}\mathbb{I}_{\ell^c}R_I^*H_{I,I}\cdot\sum_{k\ge0}H_{I,I}^k\Psi_I^*Z_I\cdot Z_I^*\mathbb{1}_I(\ell)]\| \\
&\qquad\le \|\mathbb{E}[D_{\sqrt{\pi}}^{-1}\mathbb{I}_{\ell^c}R_I^*H_{I,I}H_{I,I}^*R_I\mathbb{I}_{\ell^c}D_{\sqrt{\pi}}^{-1}\mathbb{1}_I(\ell)]\|^{\frac12}\cdot\max_{I\in\mathcal{G}}\|\sum_{k\ge0}H_{I,I}^k\Psi_I^*Z_I\|\cdot\|\mathbb{E}[Z_IZ_I^*\mathbb{1}_I(\ell)]\|^{\frac12} \\
&\qquad\le \sqrt{9\pi_\ell}\cdot\max\{\mu(\Psi),\|\Psi D_{\sqrt{\pi}}\|\}\cdot\frac{2(1+\vartheta)}{1-\vartheta}\cdot\sqrt{2\pi_\ell}\cdot\max\{\varepsilon,\|ZD_{\sqrt{\pi}}\|\} \\
&\qquad\le 15\cdot\pi_\ell\cdot\max\{\mu(\Psi),\|\Psi D_{\sqrt{\pi}}\|\}\cdot\delta.
\end{aligned}$$

Using the same tools as above with Corollary 4.20(j,f) we get for Term IIc

$$\begin{aligned}
&\|\mathbb{E}_{\mathcal{G}}[\mathbb{I}_{\ell^c}\cdot D_{\sqrt{\pi}}^{-1}R_I^*\Psi_I^*Z_IZ_I^*\Psi_I\sum_{k\ge1}H_{I,I}^kR_I\cdot e_\ell]\| \\
&\qquad= \|\mathbb{E}_{\mathcal{G}}[\mathbb{I}_{\ell^c}D_{\sqrt{\pi}}^{-1}R_I^*\Psi_I^*Z_I\mathbb{1}_I(\ell)\cdot Z_I^*\Psi_I\sum_{k\ge0}H_{I,I}^k\cdot H_{I,\ell}\mathbb{1}_I(\ell)]\| \\
&\qquad\le \|\mathbb{E}[D_{\sqrt{\pi}}^{-1}\mathbb{I}_{\ell^c}R_I^*\Psi_I^*Z_IZ_I^*\Psi_I^*R_I\mathbb{I}_{\ell^c}D_{\sqrt{\pi}}^{-1}\mathbb{1}_I(\ell)]\|^{\frac12}\cdot\max_{I\in\mathcal{G}}\|Z_I^*\Psi_I\sum_{k\ge0}H_{I,I}^k\|\cdot\|\mathbb{E}[H_{\ell,I}H_{I,\ell}\mathbb{1}_I(\ell)]\|^{\frac12} \\
&\qquad\le 3\sqrt{\pi_\ell}\cdot\max\{\varepsilon,\|ZD_{\sqrt{\pi}}\|\}\cdot\frac{2(1+\vartheta)}{1-\vartheta}\cdot\sqrt{\pi_\ell}\cdot\|\Psi D_{\sqrt{\pi}}\| \le 10\cdot\pi_\ell\cdot\|\Psi D_{\sqrt{\pi}}\|\cdot\delta.
\end{aligned}$$

To bound the norm term corresponding to IId, we again combine the bound $\|Z_I\| \le \|\Psi_I\| + \|\Phi_I\| \le 2\sqrt{1+\vartheta}$ with the bound due to the definition of $\mathcal{G}$, cp. (4.47), ensuring that $\|Z_I\|$ is close to δ on $\mathcal{G}$, and get for $\vartheta \le 1/4$

$$\max_{I\in\mathcal{G}}\|Z_IZ_I^*\| \le 2\sqrt{1+\vartheta}\cdot\sqrt{2e^2\log(320K\rho/\delta_\star)}\cdot\delta \le \sqrt{10e^2\log(320K\rho/\delta_\star)}\cdot\delta.$$

Thus by applying Theorem 4.7 and Corollary 4.20(h,f) together with $\vartheta \leq 1/4$ we have

$$\begin{aligned}
&\|\mathbb{E}_{\mathcal{G}}[\mathbb{I}_{\ell^c} \cdot D_{\sqrt{\pi}}^{-1} R_I^* \sum_{k\geq 1} H_{I,I}^k \Psi_I^* Z_I Z_I^* \Psi_I \sum_{k\geq 1} H_{I,I}^k R_I \cdot e_\ell]\| \\
&\quad = \|\mathbb{E}_{\mathcal{G}}[\mathbb{I}_{\ell^c} D_{\sqrt{\pi}}^{-1} R_I^* H_{I,I} \mathbb{1}_I(\ell) \cdot \sum_{k\geq 0} H_{I,I}^k \Psi_I^* Z_I Z_I^* \Psi_I \sum_{k\geq 0} H_{I,I}^k \cdot H_{I,\ell} \mathbb{1}_I(\ell)]\| \\
&\quad \leq \|\mathbb{E}[D_{\sqrt{\pi}}^{-1} \mathbb{I}_{\ell^c} R_I^* H_{I,I} H_{I,I}^* R_I \mathbb{I}_{\ell^c} D_{\sqrt{\pi}}^{-1} \mathbb{1}_I(\ell)]\|^{\frac{1}{2}} \\
&\qquad\qquad \cdot \max_{I\in\mathcal{G}} \|\sum_{k\geq 0} H_{I,I}^k \Psi_I^* Z_I Z_I^* \Psi_I \sum_{k\geq 0} H_{I,I}^k\| \cdot \|\mathbb{E}[H_{\ell,I} H_{I,\ell} \mathbb{1}_I(\ell)]\|^{\frac{1}{2}} \\
&\quad \leq \sqrt{9\pi_\ell} \cdot \max\{\mu(\Psi), \|\Psi D_{\sqrt{\pi}}\|\} \cdot \frac{(1+\vartheta)}{(1-\vartheta)^2} \cdot \sqrt{10e^2 \log(320K\rho/\delta_\star)} \cdot \delta \cdot \sqrt{\pi_\ell} \cdot \|\Psi D_{\sqrt{\pi}}\| \\
&\quad \leq m_{11} \cdot \pi_\ell \cdot \max\{\mu(\Psi), \|\Psi D_{\sqrt{\pi}}\|\} \cdot \delta,
\end{aligned}$$

where in the last step we used the assumptions $\|\Psi D_{\sqrt{\pi}}\| \leq C^{-1}/\sqrt{\log(K\rho/\delta_\star)}$. Plugging everything back into (4.98) and using that by Lemma 4.11 $\mathbb{P}(\mathcal{H}^c)\, 2\rho + \mathbb{P}(\mathcal{G}^c)\, \rho \leq \frac{1}{32}\delta_\star$ we get

$$\|\mathbb{E}[\hat{Y}]\| \leq \frac{1}{32}\delta_\star + m_{12}\|D_\alpha^{-2}\| \cdot \Lambda \cdot \big(\|ZD_{\sqrt{\pi}}\| \cdot (1+\delta) + \max\{\mu(\Psi), \|\Psi D_{\sqrt{\pi}}\|\} \cdot \delta\big).$$

Using again that $\|ZD_{\sqrt{\pi}}\| \leq \min\{\delta, \|\Phi D_{\sqrt{\pi}}\| + \|\Psi D_{\sqrt{\pi}}\|\}$ and our assumptions that

$$\max\{\mu(\Psi), \|\Psi D_{\sqrt{\pi}}\|, \|\Phi D_{\sqrt{\pi}}\|\} \leq \frac{\underline{\alpha}\gamma}{C\sqrt{\log(K\rho/\delta_\star)}},$$

together with the definition of $\Lambda = \frac{\underline{\alpha}\gamma}{C\sqrt{\log(K\rho/\delta_\star)}}$, we get for $C \geq 256 m_{13}$

$$\begin{aligned}
\|\mathbb{E}[\hat{Y}]\| &\leq \frac{1}{32}\delta_\star + m_{13}\|D_\alpha^{-2}\| \frac{\underline{\alpha}\gamma}{C\sqrt{\log(K\rho/\delta_\star)}} \min\left\{\frac{\underline{\alpha}\gamma}{C\sqrt{\log(K\rho/\delta_\star)}}, \delta\right\} \\
&\leq \frac{1}{32}\delta_\star + \frac{1}{2}\min\left\{\frac{\gamma^2}{128C\log(K\rho/\delta_\star)}, \frac{\gamma}{128\sqrt{\log(K\rho/\delta_\star)}} \cdot \delta\right\} = \frac{1}{2}\Delta := m.
\end{aligned}$$

As before an application of the vector Bernstein inequality for $t = \Delta/2$ with $R = \frac{3}{4}\rho$ and $m = \Delta/2$, and some simplifications yield the desired bound. ■

Remark 4.16 *Here we collect all lower bounds on C. C has to be big enough to ensure that the probabilities in* (4.55) *are small. Further also the conditions*

$$C \geq \sqrt{2} \cdot 256 \max\{m_4, m_7, m_{10}, m_{13}\}$$

have to be satisfied. We defer calculating the exact constants to those brave enough.

4.9. How to calculate expectations in the rejective sampling model

In this section we establish the probabilistic estimates used to prove Claims 1-4. In the case of supports chosen uniformly at random, equivalent to rejective sampling with uniform weights $p_i = S/K$ most estimates are trivial since we have $\mathbb{P}_S(i \in I) = S/K$ and

$$\mathbb{P}_S(\{i,j\} \subseteq I) = \frac{S(S-1)}{K(K-1)} = c \cdot \mathbb{P}_S(i \in I) \cdot \mathbb{P}_S(j \in I),$$

so for instance $\|\mathbb{E}_S[R_I^* H_{I,I} R_I]\| = \|H \odot \mathbb{E}_S[\mathbf{1}_I \mathbf{1}_I^*]\| = \frac{S}{K}\|\Psi^*\Psi - \mathbb{I}\| \leq \frac{S}{K}\|\Psi\|^2$. Unfortunately, in the rejective sampling model with non-uniform weights p_i, these estimates become much more involved. For instance the appearance probabilities $\pi(i) = \mathbb{P}_S(i \in I)$ differ from p_i and higher order appearance probabilities cannot simply be obtained as scaled product of low order appearance probabilities, i.e. there is no constant c such that $\mathbb{P}_S(i,j \in I) = c\pi(i)\pi(j)$ for all i,j. However, we can show that for well behaved p_i, the generating and appearance probabilities are close, and that also higher order appearance probabilities are close to products of lower ones, which will allow us to bound quantities such as $\|\mathbb{E}_S[\Psi_I \Psi_I^*]\|$ later on.
We first establish several inequalities for appearance probabilities in the rejective sampling model.

Theorem 4.17 *Let $\mathbb{P}_B$ be the probability measure corresponding to the Poisson sampling model (1.2) with weights $p_i < 1$ and $\mathbb{P}_S$ be the probability measure corresponding to the associated rejective sampling model with parameter S, $\mathbb{P}_S(I) = \mathbb{P}_B(I \mid |I| = S)$, as well as $\mathbb{E}_S$ the expectation with respect to $\mathbb{P}_S$. Denote by π_S the vector of first order inclusion probabilities of level S, meaning $\pi_S(i) = \mathbb{P}_S(i \in I)$ or equivalently $\pi_S = \mathbb{E}_S(\mathbf{1}_I)$. We have*

$$(1 - \|p\|_\infty) \cdot p_i \leq \pi_S(i) \leq 2 \cdot p_i, \quad \text{if} \quad \textstyle\sum_k p_k = S, \tag{a}$$

$$\pi_{S-1}(i) \leq \pi_S(i), \tag{b}$$

$$\mathbb{P}_S(i \in I, L \subseteq I) \leq \pi_S(i) \cdot \mathbb{P}_S(L \subseteq I), \quad \text{if} \quad |L| < S, i \notin L, \tag{c}$$

$$[1 - \pi_{S-1}(i)] \cdot \mathbb{P}_S(\{i,j\} \subseteq I) = \pi_S(i) \cdot [\pi_{S-1}(j) - \mathbb{P}_{S-1}(\{i,j\} \subseteq I)], \quad \text{if} \quad i \neq j. \tag{d}$$

Further, defining for $L \subseteq [K]$ with $|L| < S$ the set $\mathcal{L} = \{I \subseteq [K] : L \subseteq I\}$, we have

$$\mathbb{E}_S\big[\mathbf{1}_{I \setminus L} \mathbf{1}_{I \setminus L}^* \cdot \mathbb{1}_{\mathcal{L}}(I)\big] \cdot \prod_{\ell \in L} [1 - \pi_S(\ell)] \preceq \mathbb{E}_{S-|L|}[\mathbf{1}_I \mathbf{1}_I^*] \cdot \prod_{\ell \in L} \pi_S(\ell). \tag{e}$$

Proof **(a)** We first show that $(1 - \|p\|_\infty)p_i \leq \pi_S(i)$. By definition, we have

$$\pi_S(i) = \mathbb{P}_B(i \in I \mid |I| = S) = \frac{\mathbb{P}_B(\{i \in I\} \cap \{|I| = S\})}{\mathbb{P}_B(|I| = S)} = \frac{\sum_{I:|I|=S, i \in I} \mathbb{P}_B(I)}{\sum_{I:|I|=S} \mathbb{P}_B(I)} \underbrace{\sum_J \mathbb{P}_B(J)}_{=1}.$$

Since $p_i = \sum_{J: i \in J} \mathbb{P}_B(J)$ and abbreviating $c := (1 - \|p\|_\infty) \leq 1$ the desired inequality

$$c \sum_{J: i \in J} \mathbb{P}_B(J) \leq \frac{\sum_{I:|I|=S, i \in I} \mathbb{P}_B(I)}{\sum_{I:|I|=S} \mathbb{P}_B(I)} \sum_J \mathbb{P}_B(J),$$

is equivalent to

$$c \sum_{I:|I|=S} \mathbb{P}_B(I) \sum_{J: i \in J} \mathbb{P}_B(J) \leq \sum_{I:|I|=S, i \in I} \mathbb{P}_B(I) \sum_J \mathbb{P}_B(J).$$

Splitting the sums on both sides into two parts this becomes

$$c \sum_{I:|I|=S,i\notin I} \mathbb{P}_B(I) \sum_{J:i\in J} \mathbb{P}_B(J) + c \sum_{I:|I|=S,i\in I} \mathbb{P}_B(I) \sum_{J:i\in J} \mathbb{P}_B(J)$$
$$\leq \sum_{I:|I|=S,i\in I} \mathbb{P}_B(I) \sum_{J:i\in J} \mathbb{P}_B(J) + \sum_{I:|I|=S,i\in I} \mathbb{P}_B(I) \sum_{J:i\notin J} \mathbb{P}_B(J),$$

which is implied by

$$c \sum_{I:|I|=S,i\notin I} \mathbb{P}_B(I) \sum_{J:i\in J} \mathbb{P}_B(J) \leq \sum_{I:|I|=S,i\in I} \mathbb{P}_B(I) \sum_{J:i\notin J} \mathbb{P}_B(J). \tag{4.100}$$

Note that for any set I not containing the index i we have

$$\frac{p_i}{1-p_i} \cdot \mathbb{P}_B(I) = \frac{p_i}{1-p_i} \prod_{k\in I} p_k \prod_{k\notin I} (1-p_k) = \prod_{k\in I\cup\{i\}} p_k \prod_{k\notin I\cup\{i\}} (1-p_k) = \mathbb{P}_B(I\cup\{i\}).$$

Multiplying both sides in 4.100 with $p_i/(1-p_i)$ we get

$$c \sum_{I:|I|=S+1,i\in I} \mathbb{P}_B(I) \sum_{J:i\in J} \mathbb{P}_B(J) \leq \sum_{I:|I|=S,i\in I} \mathbb{P}_B(I) \sum_{J:i\in J} \mathbb{P}_B(J),$$

so it suffices to show that

$$c \sum_{I:|I|=S+1,i\in I} \mathbb{P}_B(I) \leq \sum_{I:|I|=S,i\in I} \mathbb{P}_B(I).$$

Indeed we have

$$\begin{aligned}
c \sum_{I:|I|=S+1,i\in I} \mathbb{P}_B(I) &= c \sum_{I:|I|=S+1,i\in I} \mathbb{P}_B(I) \sum_{k:k\in I,k\neq i} \frac{1}{S} \\
&= c \sum_{I:|I|=S+1,i\in I} \frac{1}{S} \sum_{k:k\in I,k\neq i} \mathbb{P}_B(I\setminus\{k\}) \frac{p_k}{1-p_k} \\
&\leq \frac{c}{S(1-\|p\|_\infty)} \sum_{\substack{(I,k):|I|=S+1,i\in I \\ k\in I,k\neq i}} \mathbb{P}_B(I\setminus\{k\}) \cdot p_k \\
&\overset{\text{scary}}{=} \frac{1}{S} \sum_{\substack{(J,k):|J|=S, \\ i\in J,k\notin J}} \mathbb{P}_B(J) \cdot p_k \\
&= \frac{1}{S} \sum_{J:|J|=S,i\in J} \mathbb{P}_B(J) \sum_{k\notin J} p_k \leq \sum_{J:|J|=S,i\in J} \mathbb{P}_B(J),
\end{aligned}$$

where we used that $\sum_{k\notin J} p_k \leq \sum_k p_k = S$.
To get $\pi_S(i) \leq 2p_i$ we define the function $f : \mathcal{P}([K]) \to \{0,1\}$ with $f(I) = \mathbb{1}_I(i)$. Since $f(I) \leq f(J)$ whenever $I \subseteq J$, applying Lemma 2.5 from Chapter 2 yields $\mathbb{P}_S(f(I)=1) \leq 2\mathbb{P}_B(f(I)=1)$. Since $f(I)=1$ simply means that $i\in I$, we get

$$\pi_S(i) = \mathbb{P}_S(i\in I) \leq 2\mathbb{P}_B(i\in I) = 2p_i,$$

which completes the proof of (a). **(a)✓**

(b) Using the definition of $\mathbb{P}_S$ we can rewrite $\pi_{S-1}(i) = \mathbb{P}_{S-1}(i \in I) \leq \mathbb{P}_S(i \in I) = \pi_S(i)$ as

$$\frac{\sum_{J:|J|=S-1} \mathbb{1}_J(i) \cdot \mathbb{P}_B(J)}{\sum_{J:|J|=S-1} \mathbb{P}_B(J)} \leq \frac{\sum_{I:|I|=S} \mathbb{1}_I(i) \cdot \mathbb{P}_B(I)}{\sum_{I:|I|=S} \mathbb{P}_B(I)},$$

which is equivalent to

$$\sum_{(I,J):|J|=S-1,|I|=S} \mathbb{1}_J(i) \cdot \mathbb{P}_B(J)\mathbb{P}_B(I) \leq \sum_{(I,J):|J|=S-1,|I|=S} \mathbb{1}_I(i) \cdot \mathbb{P}_B(J)\mathbb{P}_B(I).$$

Now the crucial step, which we will use several times also in the subsequent proofs, is to see that we can partition these sums in a special way. For a pair (I, J), by definition of the Poisson sampling model, we can write $\mathbb{P}_B(I)\mathbb{P}_B(J)$ in the following way

$$\mathbb{P}_B(I)\mathbb{P}_B(J) = \prod_{i\in I} p_i \prod_{j\notin I}(1-p_j) \prod_{i\in J} p_i \prod_{j\notin J}(1-p_j) = \prod_{i\in I\cap J} p_i^2 \prod_{i\in I\triangle J} p_i(1-p_i) \prod_{j\notin I\cup J} (1-p_j)^2,$$

where $I\triangle J$ denotes the symmetric difference of I, J. This implies that if for two pairs (I, J), (I', J') we have

$$I \cap J = I' \cap J' \quad \text{and} \quad I\triangle J = I'\triangle J' \quad \text{then} \quad \mathbb{P}_B(I)\mathbb{P}_B(J) = \mathbb{P}_B(I')\mathbb{P}_B(J').$$

This allows us to define natural partitions on the set of pairs (I, J) such that the probability $\mathbb{P}_B(I)\mathbb{P}_B(J)$ is constant on each partition. Concretely, for any integer $T \in \{1, \dots, S\}$, together with a set $A \subseteq \mathbb{K}$ with $|A| = S - T$ and a set $B \subseteq \mathbb{K} \setminus A$ with $|B| = 2T - 1$, we look at the collection of pairs (I, J) with intersection A and symmetric difference B, that is

$$\mathcal{Q}_{A,B} := \{(I, J) : I, J \subseteq \mathbb{K}, |I| = S, |J| = S - 1, I \cap J = A, I\triangle J = B\}.$$

Since each pair (I, J) with $|I| = S, |J| = S - 1$ can be *uniquely* assigned to a collection $\mathcal{Q}_{A,B}$ and $\mathbb{P}(I)\mathbb{P}(J)$ is constant for all $(I, J) \in \mathcal{Q}_{A,B}$, it is sufficient to show that

$$\sum_{(I,J)\in\mathcal{Q}_{A,B}} \mathbb{1}_J(j) \leq \sum_{(I,J)\in\mathcal{Q}_{A,B}} \mathbb{1}_I(j)$$

or equivalently that

$$|\{(I, J) \in \mathcal{Q}_{A,B} : j \in J\}| \leq |\{(I, J) \in \mathcal{Q}_{A,B} : j \in I\}. \tag{4.101}$$

If $j \in A = I \cap J$ both sides in (4.101) are equal to $|\mathcal{Q}_{A,B}|$ so the inequality trivially holds. In case $j \in B$, we distinguish between $T = 1$, meaning $|A| = S - 1$ and $T \geq 2$. For $T = 1$ the only possible configuration where $|J| = S - 1$ and $|I| = S$ is $J = A$ and $I = A \cup \{j\}$, so the left hand side is zero while the right hand side is one, again satisfying the inequality. Finally, for $T \geq 2$ we have

$$|\{(I, J) \in \mathcal{Q}_{A,B} : j \in J\}| = \binom{2T}{T-1} \leq \binom{2T}{T} = |\{(I, J) \in \mathcal{Q}_{A,B} : j \in I\}|,$$

which completes the proof of (b). **(b)✓**

(c) We define $\mathcal{L} = \{I \subseteq [K] : L \subseteq I\}$. Using this together with the definition of $\mathbb{P}_S$ we can rewrite $\mathbb{P}_S(i \in I, L \subseteq I) \leq \pi_S(i) \cdot \mathbb{P}_S(L \subseteq I)$ as

$$\frac{\sum_{I:|I|=S} \mathbb{1}_I(i) \cdot \mathbb{1}_{\mathcal{L}}(I) \cdot \mathbb{P}_B(I)}{\sum_{I:|I|=S} \mathbb{P}_B(I)} \leq \frac{\sum_{J:|J|=S} \mathbb{1}_J(i) \cdot \mathbb{P}_B(J)}{\sum_{J:|J|=S} \mathbb{P}_B(J)} \cdot \frac{\sum_{I:|I|=S} \mathbb{1}_{\mathcal{L}}(I) \cdot \mathbb{P}_B(I)}{\sum_{I:|I|=S} \mathbb{P}_B(I)},$$

which is equivalent to

$$\sum_{(I,J):|I|=|J|=S} \mathbb{1}_I(i) \cdot \mathbb{1}_{\mathcal{L}}(I) \cdot \mathbb{P}_B(I)\mathbb{P}_B(J) \leq \sum_{(I,J):|I|=|J|=S} \mathbb{1}_J(i) \cdot \mathbb{1}_{\mathcal{L}}(I) \cdot \mathbb{P}_B(I)\mathbb{P}_B(J).$$

We now use a similar decomposition as before. For $T \in \{0, \dots, S\}$, $A \subseteq \mathbb{K}$ with $|A| = S - T$ and $B \subseteq \mathbb{K} \setminus A$ with $|B| = 2T$, we again let A be the intersection and B the symmetric difference of the sets I and J respectively and for any combination A, B define

$$\mathcal{Q}_{A,B} := \{(I,J) : I, J \subseteq \mathbb{K}, |I| = |J| = S, I \cap J = A, I \triangle J = B\}.$$

Since $\mathbb{P}(I)\mathbb{P}(J)$ is constant for all $(I,J) \in \mathcal{Q}_{A,B}$ and every pair (I,J) is contained in exactly one of those sets, it is sufficient to show that

$$\sum_{(I,J)\in\mathcal{Q}_{A,B}} \mathbb{1}_I(i) \cdot \mathbb{1}_{\mathcal{L}}(I) \leq \sum_{(I,J)\in\mathcal{Q}_{A,B}} \mathbb{1}_J(i) \cdot \mathbb{1}_{\mathcal{L}}(I)$$

or equivalently that

$$|\{(I,J) \in \mathcal{Q}_{A,B} : i \in I, L \subseteq I\}| \leq |\{(I,J) \in \mathcal{Q}_{A,B} : i \in J, L \subseteq I\}|.$$

If L is not contained in $A \cup B$ there is no valid pair $(I,J) \in \mathcal{Q}_{A,B}$ and the inequality trivially holds. If $L \subseteq A \cup B$ we abbreviate $L_A = L \cap A$ and $L_B = L \cap B$. Since $L_A \subseteq A$, all pairs in $(I,J) \in \mathcal{Q}_{A,B}$ automatically satisfy $L_A \subseteq I$ so we can rewrite the inequality we want to show as

$$|\{(I,J) \in \mathcal{Q}_{A,B} : i \in I, L_B \subseteq I\}| \leq |\{(I,J) \in \mathcal{Q}_{A,B} : i \in J, L_B \subseteq I\}|. \tag{4.102}$$

In case $i \in A$ all pairs further satisfy $i \in I$ as well as $i \in J$ so both sides in (4.102) correspond to $|\{(I,J) \in \mathcal{Q}_{A,B} : L_B \subseteq I\}|$ and the inequality trivially holds. In case $i \in B \setminus L$, we need to keep track of how many elements of I are already fixed. Since we need to have $A \cup L_B \subseteq I$, in case $|A \cup L_B| = |A| + |L_B| = S$ there is no pair which additionally satisfies $i \in I$, so the left hand side in (4.102) is zero and the inequality holds. Finally, if $k = |L_B| < S - |A| = T$, we can still choose $T - k - 1$ out of the $2T - k - 1$ resp. $T - k$ out of the $2T - k - 1$ remaining elements in B to fill I resp. J and create a valid pair. Since

$$\binom{2T-k-1}{T-k-1} \leq \binom{2T-k-1}{T-k}$$

the inequality in (4.102) is again satisfied, which completes the proof of (c). **(c)✓**

(d) We want to show that $[1 - \pi_{S-1}(i)] \cdot \mathbb{P}_S(\{i,j\} \subseteq I) = \pi_S(i) \cdot [\pi_{S-1}(j) - \mathbb{P}_{S-1}(\{i,j\} \subseteq I)]$. If $\pi_{S-1}(i) < 1$ this is equivalent to $p_i < 1$, so for any set J not containing the index i we have

$$\frac{p_i}{1-p_i} \cdot \mathbb{P}_B(J) = \mathbb{P}_B(J \cup \{i\}).$$

$$\begin{aligned}
\mathbb{P}_S(\{i,j\} \subseteq I) &= \frac{\sum_{I:|I|=S} \mathbb{1}_I(i)\mathbb{1}_I(j) \cdot \mathbb{P}_B(I)}{\sum_{I:|I|=S} \mathbb{P}_B(I)} \cdot \frac{\sum_{I:|I|=S,i\in I} \mathbb{P}_B(I)}{\sum_{I:|I|=S,i\in I} \mathbb{P}_B(I)} \\
&= \frac{\sum_{I:|I|=S,i\in I} \mathbb{1}_I(j) \cdot \mathbb{P}_B(I)}{\sum_{|I|=S,i\in I} \mathbb{P}_B(I)} \cdot \pi_S(i) \\
&= \frac{p_i}{1-p_i} \cdot \frac{1-p_i}{p_i} \cdot \frac{\sum_{J:|J|=S-1,i\notin J} \mathbb{1}_J(j) \cdot \mathbb{P}_B(J)}{\sum_{J:|J|=S-1,i\notin J} \mathbb{P}_B(J)} \cdot \pi_S(i) \cdot \frac{\sum_{J:|J|=S-1} \mathbb{P}_B(J)}{\sum_{J:|J|=S-1} \mathbb{P}_B(J)} \\
&= \pi_S(i) \cdot \frac{\sum_{J:|J|=S-1,i\notin J} \mathbb{1}_J(j) \cdot \mathbb{P}_B(J)}{\sum_{J:|J|=S-1} \mathbb{P}_B(J)} \cdot \frac{\sum_{J:|J|=S-1} \mathbb{P}_B(J)}{\sum_{J:|J|=S-1,i\notin J} \mathbb{P}_B(J)}.
\end{aligned}$$

Further rewriting the fractions in the expression above yields

$$\frac{\sum_{J:|J|=S-1,i\notin J}\mathbb{1}_J(j)\cdot\mathbb{P}_B(J)}{\sum_{J:|J|=S-1}\mathbb{P}_B(J)}=\underbrace{\frac{\sum_{J:|J|=S-1}\mathbb{1}_J(j)\cdot\mathbb{P}_B(I)}{\sum_{J:|J|=S-1}\mathbb{P}_B(J)}}_{\pi_{S-1}(j)}-\underbrace{\frac{\sum_{J:|J|=S-1}\mathbb{1}_J(i)\mathbb{1}_J(j)\cdot\mathbb{P}_B(J)}{\sum_{J:|J|=S-1}\mathbb{P}_B(J)}}_{\mathbb{P}_{S-1}(\{i,j\in I\})}$$

as well as

$$\frac{\sum_{J:|J|=S-1}\mathbb{P}_B(J)}{\sum_{J:|J|=S-1,i\notin J}\mathbb{P}_B(J)}=\frac{\sum_{J:|J|=S-1}\mathbb{P}_B(J)}{\sum_{J:|J|=S-1}\mathbb{P}_B(J)-\sum_{J:|J|=S-1,i\in J}\mathbb{P}_B(J)}=\frac{1}{1-\pi_{S-1}(i)},$$

which completes the proof of (d). **(d)✓**

(e) We will prove the statement by induction. Let $\hat{L}$ be a set of size $T\leq S-2$ and $\hat{\mathcal{L}}=\{I\subseteq[K]:\hat{L}\subseteq I\}$. We first show that for $k\notin\hat{L}$ and $L=\hat{L}\cup\{k\}$ we have

$$(1-\pi_S(k))\cdot\mathbb{E}_S\big[\mathbf{1}_{I\setminus L}\mathbf{1}^*_{I\setminus L}\cdot\mathbb{1}_{\mathcal{L}}(I)\big]\preceq\pi_S(k)\cdot\mathbb{E}_{S-1}\big[\mathbf{1}_{I\setminus\hat{L}}\mathbf{1}^*_{I\setminus\hat{L}}\cdot\mathbb{1}_{\hat{\mathcal{L}}}(I)\big]. \tag{4.103}$$

Note that if $\pi_S(k)=1$ the inequality is trivially true. If on the other hand $\pi_S(k)<1$ this is equivalent to $p_k<1$, so for any set J not containing the index k we have

$$\frac{p_k}{1-p_k}\cdot\mathbb{P}_B(J)=\mathbb{P}_B(J\cup\{k\}).$$

Thus expanding the expectation we get

$$\begin{aligned}
\mathbb{E}_S\big[\mathbf{1}_{I\setminus L}\mathbf{1}^*_{I\setminus L}\cdot\mathbb{1}_{\mathcal{L}}(I)\big]&=\frac{\sum_{I:|I|=S,L\subseteq I}\mathbb{P}_B(I)(\mathbf{1}_{I\setminus L}\mathbf{1}^*_{I\setminus L})}{\sum_{I:|I|=S}\mathbb{P}_B(I)}\cdot\frac{\sum_{I:|I|=S,k\in I}\mathbb{P}_B(I)}{\sum_{I:|I|=S,k\in I}\mathbb{P}_B(I)}\\
&=\frac{\sum_{I:|I|=S,L\subseteq I}\mathbb{P}_B(I)(\mathbf{1}_{I\setminus L}\mathbf{1}^*_{I\setminus L})}{\sum_{I:|I|=S,k\in I}\mathbb{P}_B(I)}\cdot\frac{\sum_{I:|I|=S,k\in I}\mathbb{P}_B(I)}{\sum_{I:|I|=S}\mathbb{P}_B(I)}\\
&=\frac{\sum_{J:|J|=S-1,k\notin J,\hat{L}\subseteq J}\mathbb{P}_B(J)(\mathbf{1}_{J\setminus\hat{L}}\mathbf{1}^*_{J\setminus\hat{L}})}{\sum_{J:|J|=S-1,k\notin J}\mathbb{P}_B(J)}\cdot\pi_S(k)\\
&\preceq\frac{\sum_{J:|J|=S-1,\hat{L}\subseteq J}\mathbb{P}_B(J)(\mathbf{1}_{J\setminus\hat{L}}\mathbf{1}^*_{J\setminus\hat{L}})}{\sum_{J:|J|=S-1,k\notin J}\mathbb{P}_B(J)}\cdot\frac{\sum_{I:|I|=S-1}\mathbb{P}_B(I)}{\sum_{I:|I|=S-1}\mathbb{P}_B(I)}\cdot\pi_S(k)\\
&=\frac{\sum_{J:|J|=S-1,\hat{L}\subseteq J}\mathbb{P}_B(J)(\mathbf{1}_{J\setminus\hat{L}}\mathbf{1}^*_{J\setminus\hat{L}})}{\sum_{I:|I|=S-1}\mathbb{P}_B(I)}\cdot\frac{\sum_{I:|I|=S-1}\mathbb{P}_B(I)}{\sum_{J:|J|=S-1,k\notin J}\mathbb{P}_B(J)}\cdot\pi_S(k)\\
&=\mathbb{E}_{S-1}\big[\mathbf{1}_{I\setminus\hat{L}}\mathbf{1}^*_{I\setminus\hat{L}}\cdot\mathbb{1}_{\hat{\mathcal{L}}}(I)\big]\cdot\frac{\sum_{I:|I|=S-1}\mathbb{P}_B(I)}{\sum_{J:|J|=S-1,k\notin J}\mathbb{P}_B(J)}\cdot\pi_S(k).
\end{aligned}$$

Now all that remains to do in order to prove (4.103) is to bound the fraction above. Writing out the expression in the denominator we get

$$\begin{aligned}
\frac{\sum_{I:|I|=S-1}\mathbb{P}_B(I)}{\sum_{J:|J|=S-1,k\notin J}\mathbb{P}_B(J)}&=\frac{\sum_{I:|I|=S-1}\mathbb{P}_B(I)}{\sum_{I:|I|=S-1}\mathbb{P}_B(I)-\sum_{I:|I|=S-1,k\in I}\mathbb{P}_B(I)}\\
&=\frac{1}{1-\mathbb{P}_{S-1}(k\in I)}\overset{\text{(b)}}{\leq}\frac{1}{1-\mathbb{P}_S(k\in I)}=\frac{1}{1-\pi_S(k)}.
\end{aligned}$$

By induction and using the bound from (b) that $\pi_{S-1}(k) \leq \pi_S(k)$ we finally get

$$\mathbb{E}_S\big[\mathbf{1}_{I\setminus L}\mathbf{1}_{I\setminus L}^* \cdot \mathbb{1}_{\mathcal{L}}(I)\big] \prod_{\ell\in L}(1-\pi_S(\ell)) \preceq \mathbb{E}_{S-|L|}[\mathbf{1}_I\mathbf{1}_I^*] \cdot \prod_{\ell\in L}\pi_S(\ell).$$

which completes the proof of (e) and thus the theorem. **(e)✓**

■

We next note that several of the quantities we want to bound can be conveniently rewritten using the entry-wise product between matrices, ie. $(A \odot B)_{ij} = A_{ij} \cdot B_{ij}$, also known as Hadamard product. For instance the zero padded $K \times K$ version of the $S \times S$ matrix $A_{I,I}$ can be written as $R_I^* A_{I,I} R_I = A \odot (\mathbf{1}_I\mathbf{1}_I^*)$. Therefore the following inequality will be an essential tool.

Theorem 4.18 (Hadamard Product Matrix Norm Inequality) *Let A and B be two square matrices of the same dimension. If A is positive-semidefinite, then*

$$\|A \odot B\|_{2,2} \leq \|A\|_{\infty,1}\|B\|_{2,2}.$$

Proof The matrix

$$\begin{pmatrix} \|B\|_{2,2}(\mathbb{I}\odot A) & A\odot B \\ (A\odot B)^* & \|B\|_{2,2}(\mathbb{I}\odot A)\end{pmatrix} = \begin{pmatrix} A & A \\ A & A\end{pmatrix} \odot \begin{pmatrix} \|B\|_{2,2}\cdot\mathbb{I} & B \\ B^* & \|B\|_{2,2}\cdot\mathbb{I}\end{pmatrix}$$

is psd, since the right hand side of the equation is a Hadamard product of two psd matrices which is by the Schur product Theorem also psd. By Theorem 7.7.9 in [44] there thus exists a contraction C, meaning $\|C\|_{2,2} \leq 1$, such that

$$A \odot B = \|B\|_{2,2}(\mathbb{I}\odot A)^{1/2} C (\mathbb{I}\odot A)^{1/2}.$$

Hence

$$\|A\odot B\|_{2,2} \leq \|(\mathbb{I}\odot A)^{1/2}\|^2 \|B\|_{2,2} \leq \|\mathbb{I}\odot A\|_{2,2}\|B\|_{2,2} \leq \|A\|_{\infty,1}\|B\|_{2,2}. \tag{4.104}$$

■

We can now provide a bound for terms of the form $\|\mathbb{E}_S[R_I^* A_{I,I} R_I]\| = \|A \odot \mathbb{E}_S[\mathbf{1}_I\mathbf{1}_I^*]\|$, which is the key result we need for most estimates used in the proofs of Claims 1-4.

Theorem 4.19 *Let $\mathbb{E}_S$ be the expectation according to the rejective sampling probability $\mathbb{P}_S$ and $\pi_S \in \mathbb{R}^K$ be the first order inclusion probabilities of level S. If $\|\pi_S\|_\infty < 1$ then for any $K \times K$ matrix A we have*

$$\|A \odot \mathbb{E}_S[\mathbf{1}_I\mathbf{1}_I^*]\| \leq \frac{1+\|\pi_S\|_\infty}{(1-\|\pi_S\|_\infty)^2} \cdot \|D_{\pi_S}[A-\operatorname{diag}(A)]D_{\pi_S}\| + \|\operatorname{diag}(A)D_{\pi_S}\|.$$

Proof We first note that since $\mathbb{E}_S[\mathbf{1}_I\mathbf{1}_I^*]$ has π_S on the diagonal, a simple application of the triangle inequality yields

$$\begin{aligned}\|A \odot \mathbb{E}_S[\mathbf{1}_I\mathbf{1}_I^*]\| &\leq \|(A-\operatorname{diag}(A)) \odot \mathbb{E}_S[\mathbf{1}_I\mathbf{1}_I^*]\| + \|\operatorname{diag}(A) \odot \mathbb{E}_S[\mathbf{1}_I\mathbf{1}_I^*]\| \\ &= \|(A-\operatorname{diag}(A)) \odot \mathbb{E}_S[\mathbf{1}_I\mathbf{1}_I^*]\| + \|\operatorname{diag}(A)D_{\pi_S}\|,\end{aligned}$$

which already proves the theorem for $S = 1$, where all off-diagonal entries of $\mathbb{E}_S[\mathbf{1}_I\mathbf{1}_I^*]$ are zero, meaning the first norm term is zero. In case $S \geq 2$ it remains to show that for $\mathit{И} = A - \operatorname{diag}(A)$ we have $\|\mathit{И} \odot \mathbb{E}_S[\mathbf{1}_I\mathbf{1}_I^*]\| \leq c \cdot \|D_{\pi_S}\mathit{И}D_{\pi_S}\|$ with constant c as above. For two vectors v, w with entries in $(0,1)$ and function $f : \mathbb{R}^2 \to \mathbb{R}$, let $\operatorname{diag}(f(v,w))$ be the diagonal matrix with i-th diagonal entry $f(v(i), w(i))$. From Theorem 4.17 we know that

$$\begin{aligned}
(\mathit{И} \odot \mathbb{E}_S[\mathbf{1}_I\mathbf{1}_I^*])_{ij} &= \mathit{И}_{ij} \cdot \mathbb{P}_S(\{i,j\} \subseteq I) \\
&= \frac{\pi_S(i)}{1-\pi_{S-1}(i)} \cdot A_{ij} \cdot [\pi_{S-1}(j) - \mathbb{P}_{S-1}(\{i,j\} \subseteq I)] \\
&= \left(\operatorname{diag}\left(\tfrac{\pi_S}{1-\pi_{S-1}}\right) \cdot \mathit{И} \cdot \operatorname{diag}(\pi_{S-1}) - \operatorname{diag}\left(\tfrac{\pi_S}{1-\pi_{S-1}}\right) \cdot \mathit{И} \odot \mathbb{E}_{S-1}[\mathbf{1}_I\mathbf{1}_I^*]\right)_{ij} \\
&= \left(\operatorname{diag}(\pi_{S-1}) \cdot \mathit{И} \cdot \operatorname{diag}\left(\tfrac{\pi_S}{1-\pi_{S-1}}\right) - \mathit{И} \cdot \operatorname{diag}\left(\tfrac{\pi_S}{1-\pi_{S-1}}\right) \odot \mathbb{E}_{S-1}[\mathbf{1}_I\mathbf{1}_I^*]\right)_{ij},
\end{aligned}$$

where for the last equality we have used the symmetry of $\mathbb{P}_S(\{i,j\} \subseteq I)$ in i, j. Using the relation between inclusion probabilities π_{S-1} and π_S from Theorem 4.17(b) we further get

$$\begin{aligned}
\|\mathit{И} \odot \mathbb{E}_S[\mathbf{1}_I\mathbf{1}_I^*]\| &\leq \|\operatorname{diag}\left(\tfrac{\pi_S}{1-\pi_{S-1}}\right) \cdot \mathit{И} \cdot \operatorname{diag}(\pi_{S-1})\| + \|\operatorname{diag}\left(\tfrac{\pi_S}{1-\pi_{S-1}}\right) \cdot \mathit{И} \odot \mathbb{E}_{S-1}[\mathbf{1}_I\mathbf{1}_I^*]\| \\
&\leq \|\operatorname{diag}\left(\tfrac{1}{1-\pi_{S-1}}\right)\| \cdot \|\operatorname{diag}(\pi_S) \cdot \mathit{И} \cdot \operatorname{diag}(\pi_S)\| \cdot \|\operatorname{diag}(\tfrac{\pi_{S-1}}{\pi_S})\| \\
&\qquad + \|\operatorname{diag}\left(\tfrac{1}{1-\pi_{S-1}}\right)\| \cdot \|\operatorname{diag}(\pi_S) \cdot \mathit{И} \odot \mathbb{E}_{S-1}[\mathbf{1}_I\mathbf{1}_I^*]\| \\
&\leq (1 - \|\pi_S\|_\infty)^{-1} \cdot (\|D_{\pi_S}\mathit{И}D_{\pi_S}\| + \|D_{\pi_S}\mathit{И} \odot \mathbb{E}_{S-1}[\mathbf{1}_I\mathbf{1}_I^*]\|)
\end{aligned}$$

as well as

$$\|\mathit{И} \odot \mathbb{E}_S[\mathbf{1}_I\mathbf{1}_I^*]\| \leq (1 - \|\pi_S\|_\infty)^{-1} \cdot (\|D_{\pi_S}\mathit{И}D_{\pi_S}\| + \|\mathit{И}D_{\pi_S} \odot \mathbb{E}_{S-1}[\mathbf{1}_I\mathbf{1}_I^*]\|)\,.$$

For $S = 2$, the matrix $\mathbb{E}_{S-1}[\mathbf{1}_I\mathbf{1}_I^*]$ is again a diagonal matrix, meaning the second norm term vanishes and we are done. For $S > 2$, observe that

$$(\mathit{И}D_{\pi_S} \odot \mathbb{E}_{S-1}[\mathbf{1}_I\mathbf{1}_I^*])^* = D_{\pi_S}\mathit{И}^* \odot \mathbb{E}_{S-1}[\mathbf{1}_I\mathbf{1}_I^*]$$

so we simply apply the inequality above to $\bar{\mathit{И}} \odot \mathbb{E}_{S-1}[\mathbf{1}_I\mathbf{1}_I^*]$ with $\bar{\mathit{И}} = D_{\pi_S}\mathit{И}^*$, leading to

$$\begin{aligned}
\|\mathit{И}D_{\pi_S} \odot \mathbb{E}_{S-1}[\mathbf{1}_I\mathbf{1}_I^*]\| &= \|D_{\pi_S}\mathit{И}^* \odot \mathbb{E}_{S-1}[\mathbf{1}_I\mathbf{1}_I^*]\| \\
&\leq (1 - \|\pi_{S-1}\|_\infty)^{-1} \cdot \left(\|D_{\pi_{S-1}}D_{\pi_S}\mathit{И}^*D_{\pi_{S-1}}\| + \|D_{\pi_S}\mathit{И}^*D_{\pi_{S-1}} \odot \mathbb{E}_{S-2}[\mathbf{1}_I\mathbf{1}_I^*]\|\right) \\
&\leq (1 - \|\pi_S\|_\infty)^{-1} \cdot (\|\pi_S\|_\infty \cdot \|D_{\pi_S}\mathit{И}D_{\pi_S}\| + \|D_{\pi_S}\mathit{И}D_{\pi_S} \odot \mathbb{E}_{S-2}[\mathbf{1}_I\mathbf{1}_I^*]\|)\,.
\end{aligned}$$

Combining the estimates we get

$$\|\mathit{И} \odot \mathbb{E}_S[\mathbf{1}_I\mathbf{1}_I^*]\| \leq (1 - \|\pi_S\|_\infty)^{-2} \cdot (\|D_{\pi_S}\mathit{И}D_{\pi_S}\| + \|D_{\pi_S}\mathit{И}D_{\pi_S} \odot \mathbb{E}_{S-2}[\mathbf{1}_I\mathbf{1}_I^*]\|)\,.$$

The final result follows from the fact that for a positiv semi-definite matrix P and any matrix B we have $\|B \odot P\| \leq \|B\| \cdot \max_{ij} |P_{ij}|$, Theorem 4.18, and that due to Theorem 4.17(c) with $L = \{k\}$ and 4.17(b) all entries of $\mathbb{E}_{S-2}[\mathbf{1}_I\mathbf{1}_I^*]$ are smaller than $\|\pi_S\|_\infty$. ■

With the last result in hand we can finally prove the following corollary, which collects the inequalities used in the proofs Claims 1-4.

Corollary 4.20 *Denote by $\mathbb{E}_S$ the expectation according to the rejective sampling probability with level S and by $\pi \in \mathbb{R}^K$ the first order inclusion probabilities of level S and let R_I be the restriction matrix to the index set I, meaning $A_I = AR_I^*$. If $\|\pi\|_\infty \leq 1/3$, we have for any matrix $И \in \mathbb{R}^{K\times K}$ with zero diagonal,*

$$\|\mathbb{E}[D_{\sqrt{\pi}}^{-1} R_I^* R_I D_{\sqrt{\pi}}^{-1}]\| \leq 1 \tag{a}$$
$$\|\mathbb{E}[D_{\sqrt{\pi}}^{-1} R_I^* И_{I,I} R_I D_{\sqrt{\pi}}^{-1}]\| \leq 3 \cdot \|D_{\sqrt{\pi}} И D_{\sqrt{\pi}}\| \tag{b}$$
$$\|\mathbb{E}[D_{\sqrt{\pi}}^{-1} R_I^* И_{I,I} И_{I,I}^* R_I D_{\sqrt{\pi}}^{-1}]\| \leq \tfrac{9}{2} \cdot \|D_{\sqrt{\pi}} И D_{\sqrt{\pi}}\|^2 + \tfrac{3}{2} \cdot \max_k \|e_k^* И D_{\sqrt{\pi}}\|^2 \tag{c}$$

Further, let $\Psi \in \mathbb{R}^{d\times K}$ be a dictionary and $H = \Psi^*\Psi - \mathbb{I}_K$ the associated hollow Gram matrix. For any $d \times K$ matrix $V = (v_1, \ldots, v_K)$ and any subset $\mathcal{G}$ be of all supports of size S, meaning $\mathcal{G} \subseteq \{I : |I| = S\}$, we have

$$\|\mathbb{E}[D_{\sqrt{\pi}}^{-1} \mathbb{I}_{\ell^c} R_I^* V_I^* \cdot \mathbb{1}_I(\ell) \cdot \mathbb{1}_{\mathcal{G}}(I)]\| \leq \pi_\ell \cdot \|V D_{\sqrt{\pi}}\| \tag{d}$$
$$\|\mathbb{E}[\Psi_I H_{I,\ell} \cdot \mathbb{1}_I(\ell) \cdot \mathbb{1}_{\mathcal{G}}(I)]\| \leq \pi_\ell \cdot \|\Psi D_{\sqrt{\pi}}\|^2 \tag{e}$$
$$\|\mathbb{E}[(H_{I,\ell})^* H_{I,\ell} \cdot \mathbb{1}_I(\ell)]\| \leq \pi_\ell \cdot \|\Psi D_{\sqrt{\pi}}\|^2. \tag{f}$$
$$\|\mathbb{E}[V_I V_I^* \cdot \mathbb{1}_I(\ell)]\| \leq \pi_\ell \cdot \big(\|V D_{\sqrt{\pi}}\|^2 + \|v_\ell\|^2\big) \tag{g}$$

Finally, if Ψ satisfies $\mu(\Psi) \leq 1/8$ and $\|\Psi D_{\sqrt{\pi}}\| \leq 1/8$ then for any $d \times K$ matrix $Z = (z_1, \ldots, z_K)$ with $\|z_k\| = \varepsilon_k \leq \varepsilon \leq \sqrt{2}$ and $|\langle z_k, \psi_k\rangle| = \varepsilon_k^2/2$ we have

$$\|\mathbb{E}[D_{\sqrt{\pi}}^{-1} \mathbb{I}_{\ell^c} R_I^* H_{I,I} H_{I,I}^* R_I \mathbb{I}_{\ell^c} D_{\sqrt{\pi}}^{-1} \cdot \mathbb{1}_I(\ell)]\| \leq 9 \cdot \pi_\ell \cdot \max\{\mu(\Psi), \|\Psi D_{\sqrt{\pi}}\|\}^2, \tag{h}$$
$$\|\mathbb{E}[\Psi_I H_{I,I} H_{I,I}^* \Psi_I^* \cdot \mathbb{1}_I(\ell)]\| \leq 2 \cdot \pi_\ell \cdot \|\Psi D_{\sqrt{\pi}}\|^2, \tag{i}$$
$$\|\mathbb{E}[D_{\sqrt{\pi}}^{-1} \mathbb{I}_{\ell^c} R_I^* \Psi_I^* Z_I Z_I^* \Psi_I R_I \mathbb{I}_{\ell^c} D_{\sqrt{\pi}}^{-1} \cdot \mathbb{1}_I(\ell)]\| \leq 9 \cdot \pi_\ell \cdot \max\{\varepsilon, \|Z D_{\sqrt{\pi}}\|\}^2, \tag{j}$$
$$\|\mathbb{E}[D_{\sqrt{\pi}}^{-1} \mathbb{I}_{\ell^c} R_I^* Z_I^* \Psi_I \Psi_I^* Z_I R_I \mathbb{I}_{\ell^c} D_{\sqrt{\pi}}^{-1} \cdot \mathbb{1}_I(\ell)]\| \leq 12 \cdot \pi_\ell \cdot \max\{\varepsilon, \|Z D_{\sqrt{\pi}}\|\}^2. \tag{k}$$

Proof

(a/b) Using the identities $A_{I,I} = R_I A R_I^*$ and $R_I^* R_I = \operatorname{diag}(\mathbf{1}_I)$, we can rewrite for a general matrix A and a diagonal matrix D

$$\begin{aligned} D R_I^* A_{I,I} R_I D &= D R_I^* R_I A R_I^* R_I D = D \operatorname{diag}(\mathbf{1}_I) A \operatorname{diag}(\mathbf{1}_I) D \\ &= \operatorname{diag}(\mathbf{1}_I) D A D \operatorname{diag}(\mathbf{1}_I) = (DAD) \odot (\mathbf{1}_I \mathbf{1}_I^*). \end{aligned}$$

Using Theorem 4.19 and $\|\pi\|_\infty \leq 1/3$ we therefore get

$$\begin{aligned} \|\mathbb{E}[D_{\sqrt{\pi}}^{-1} R_I^* A_{I,I} R_I D_{\sqrt{\pi}}^{-1}]\| &= \|\mathbb{E}[(D_{\sqrt{\pi}}^{-1} A D_{\sqrt{\pi}}^{-1}) \odot (\mathbf{1}_I \mathbf{1}_I^*)]\| \\ &= \|(D_{\sqrt{\pi}}^{-1} A D_{\sqrt{\pi}}^{-1}) \odot \mathbb{E}[\mathbf{1}_I \mathbf{1}_I^*]\| \\ &\leq 3\|D_\pi D_{\sqrt{\pi}}^{-1} [A - \operatorname{diag}(A)] D_{\sqrt{\pi}}^{-1} D_\pi\| + \|D_{\sqrt{\pi}}^{-1} \operatorname{diag}(A) D_{\sqrt{\pi}}^{-1} D_\pi\| \\ &= 3\|D_{\sqrt{\pi}} [A - \operatorname{diag}(A)] D_{\sqrt{\pi}}\| + \|\operatorname{diag}(A)\|. \end{aligned}$$

Setting $A = I$ resp. $A = И$ yields the inequalities in (a) and (b). **(a/b)✓**

(c) We again rewrite the expression, whose expectation we need to estimate.

$$\begin{aligned} R_I^* И_{I,I} И_{I,I}^* R_I &= R_I^* R_I \cdot И \cdot R_I^* R_I \cdot И^* \cdot R_I^* R_I \\ &= \operatorname{diag}(\mathbf{1}_I) \cdot И \cdot \operatorname{diag}(\mathbf{1}_I) \cdot И^* \cdot \operatorname{diag}(\mathbf{1}_I) \\ &= [И \cdot \operatorname{diag}(\mathbf{1}_I) \cdot И^*] \odot (\mathbf{1}_I \mathbf{1}_I^*) \\ &= \big(\sum_{k\in I} И_k И_k^*\big) \odot (\mathbf{1}_I \mathbf{1}_I^*) = \sum_{k\in I} (И_k И_k^*) \odot (\mathbf{1}_I \mathbf{1}_I^*). \end{aligned}$$

Since the k-th entry of $\mathit{И}_k$ and therefore both the k-th row and k-th column of $\mathit{И}_k\mathit{И}_k^*$ are zero, we have $(\mathit{И}_k\mathit{И}_k^*)\odot(\mathbf{1}_I\mathbf{1}_I^*) = (\mathit{И}_k\mathit{И}_k^*)\odot(\mathbf{1}_{I\setminus\{k\}}\mathbf{1}_{I\setminus\{k\}}^*)$, yielding

$$\begin{aligned} R_I^*\mathit{И}_{I,I}\mathit{И}_{I,I}^*R_I &= \sum_{k\in I}(\mathit{И}_k\mathit{И}_k^*)\odot(\mathbf{1}_{I\setminus\{k\}}\mathbf{1}_{I\setminus\{k\}}^*) \\ &= \sum_k(\mathbb{1}_I(k)\cdot\mathit{И}_k\mathit{И}_k^*)\odot(\mathbf{1}_{I\setminus\{k\}}\mathbf{1}_{I\setminus\{k\}}^*) \\ &= \sum_k(\mathit{И}_k\mathit{И}_k^*)\odot(\mathbf{1}_{I\setminus\{k\}}\mathbf{1}_{I\setminus\{k\}}^*\cdot\mathbb{1}_I(k)). \end{aligned} \tag{4.105}$$

Using the Schur Product Theorem, which says that for p.s.d matrices $A, P, \bar{P}$, with $P_{ij}, \bar{P}_{ij} \geq 0$ and $P \preceq \bar{P}$ we have $A\odot P \preceq A\odot\bar{P}$, together with Theorem 4.17(e) further leads to

$$\begin{aligned} \mathbb{E}_S\left[R_I^*\mathit{И}_{I,I}\mathit{И}_{I,I}^*R_I\right] &= \sum_k(\mathit{И}_k\mathit{И}_k^*)\odot\mathbb{E}_S\left[\mathbf{1}_{I\setminus\{k\}}\mathbf{1}_{I\setminus\{k\}}^*\cdot\mathbb{1}_I(k)\right] \\ &\preceq \sum_k(\mathit{И}_k\tfrac{\pi_S(k)}{1-\pi_S(k)}\mathit{И}_k^*)\odot\mathbb{E}_{S-1}[\mathbf{1}_I\mathbf{1}_I^*] \\ &= \big(\sum_k\mathit{И}_k\tfrac{\pi_S(k)}{1-\pi_S(k)}\mathit{И}_k^*\big)\odot\mathbb{E}_{S-1}[\mathbf{1}_I\mathbf{1}_I^*] \\ &= (\mathit{И}\operatorname{diag}(\tfrac{\pi_S}{1-\pi_S})\mathit{И}^*)\odot\mathbb{E}_{S-1}[\mathbf{1}_I\mathbf{1}_I^*]. \end{aligned}$$

Abbreviating $M := \mathit{И}\operatorname{diag}(\frac{\pi_S}{1-\pi_S})\mathit{И}^*$, and applying Theorem 4.19 and Theorem 4.17(b) we get

$$\begin{aligned} \|\mathbb{E}_S\left[D_{\sqrt{\pi}}^{-1}R_I^*\mathit{И}_{I,I}\mathit{И}_{I,I}^*R_ID_{\sqrt{\pi}}^{-1}\right]\| &= \|D_{\sqrt{\pi}}^{-1}\,\mathbb{E}_S\left[R_I^*\mathit{И}_{I,I}\mathit{И}_{I,I}^*R_I\right]D_{\sqrt{\pi}}^{-1}\| \\ &\leq \|(D_{\sqrt{\pi}}^{-1}MD_{\sqrt{\pi}}^{-1})\odot\mathbb{E}_{S-1}[\mathbf{1}_I\mathbf{1}_I^*]\| \\ &\leq 3\|D_{\pi_{S-1}}[D_{\sqrt{\pi}}^{-1}MD_{\sqrt{\pi}}^{-1}-\operatorname{diag}(D_{\sqrt{\pi}}^{-1}MD_{\sqrt{\pi}}^{-1})]D_{\pi_{S-1}}\| \\ &\qquad + \|\operatorname{diag}(D_{\sqrt{\pi}}^{-1}MD_{\sqrt{\pi}}^{-1})D_{\pi_{S-1}}\| \\ &\leq 3\|D_{\sqrt{\pi}}MD_{\sqrt{\pi}}-\operatorname{diag}(D_{\sqrt{\pi}}MD_{\sqrt{\pi}})\| + \|\operatorname{diag}(M)\| \\ &\leq 3\|D_{\sqrt{\pi}}MD_{\sqrt{\pi}}\| + \|\operatorname{diag}(M)\|, \end{aligned}$$

where in last inequality we have used that $D_{\sqrt{\pi}}MD_{\sqrt{\pi}}$ is positive semidefinite. Combining the inequality above with the bounds

$$\begin{aligned} \|D_{\sqrt{\pi}}MD_{\sqrt{\pi}}\| &= \|D_{\sqrt{\pi}}\mathit{И}D_{\sqrt{\pi}}\operatorname{diag}(\tfrac{1}{1-\pi})D_{\sqrt{\pi}}\mathit{И}^*D_{\sqrt{\pi}}\| \leq (1-\|\pi\|_\infty)^{-1}\|D_{\sqrt{\pi}}\mathit{И}D_{\sqrt{\pi}}\|^2, \\ \|\operatorname{diag}(M)\| &= \max_k e_k^*\mathit{И}D_{\sqrt{\pi}}\operatorname{diag}(\tfrac{1}{1-\pi})D_{\sqrt{\pi}}\mathit{И}^*e_k \leq (1-\|\pi\|_\infty)^{-1}\max_k\|e_k^*\mathit{И}D_{\sqrt{\pi}}\|^2. \end{aligned}$$

leads to (c). **(c)✓**

(d-g) Using the identity $\mathbb{1}_{\ell^c}R_I^*V_I^* = \mathbb{1}_{\ell^c}R_I^*R_IV^* = \operatorname{diag}(\mathbf{1}_{I\setminus\{\ell\}})V^*$ we get

$$\begin{aligned} \|\mathbb{E}\left[D_{\sqrt{\pi}}^{-1}\mathbb{1}_{\ell^c}R_I^*V_I^*\cdot\mathbb{1}_I(\ell)\mathbb{1}_{\mathcal{G}}(I)\right]\| &= \|D_{\sqrt{\pi}}^{-1}\mathbb{E}\left[\mathbb{1}_{\mathcal{G}}(I)\mathbb{1}_I(\ell)\cdot\operatorname{diag}(\mathbf{1}_{I\setminus\{\ell\}})\right]D_{\sqrt{\pi}}^{-1}D_{\sqrt{\pi}}V^*\| \\ &\leq \|D_{\sqrt{\pi}}^{-1}\mathbb{E}\left[\mathbb{1}_I(\ell)\cdot\operatorname{diag}(\mathbf{1}_{I\setminus\{\ell\}})\right]D_{\sqrt{\pi}}^{-1}\|\cdot\|D_{\sqrt{\pi}}V^*\| \\ &= \max_{k\neq\ell}\left(\mathbb{E}\left[\mathbb{1}_I(\ell)\mathbb{1}_I(k)\right]\cdot\pi_k^{-1}\right)\cdot\|VD_{\sqrt{\pi}}\| \\ &\leq \pi_\ell\cdot\|VD_{\sqrt{\pi}}\|, \end{aligned}$$

where in the last inequality $\mathbb{E}\left[\mathbb{1}_I(\ell)\mathbb{1}_I(k)\right] \leq \pi_\ell\pi_k$ from Theorem 4.17(c) was used.
Next observe that H has a zero diagonal, so for any $A_I = AR_I^*$ we have

$$A_IH_{I,\ell} = AR_I^*\cdot R_IHe_\ell = A\operatorname{diag}(\mathbf{1}_I)H_\ell = A\operatorname{diag}(\mathbf{1}_{I\setminus\{\ell\}})H_\ell,$$

and the same argument as above leads to

$$\begin{aligned}\|\mathbb{E}[A_I H_{I,\ell}\cdot \mathbb{1}_I(\ell)\mathbb{1}_{\mathcal{G}}(I)]\| &= \|A\,\mathbb{E}[\mathbb{1}_{\mathcal{G}}(I)\mathbb{1}_I(\ell)\cdot \operatorname{diag}(\mathbf{1}_{I\setminus\{\ell\}})]H_\ell\|\\ &\le \|AD_{\sqrt{\pi}}\|\cdot\|D_{\sqrt{\pi}}^{-1}\mathbb{E}[\mathbb{1}_I(\ell)\cdot\operatorname{diag}(\mathbf{1}_{I\setminus\{\ell\}})]D_{\sqrt{\pi}}^{-1}\|\cdot\|D_{\sqrt{\pi}}H_\ell\|\\ &\le \pi_\ell\cdot\|AD_{\sqrt{\pi}}\|\cdot\|D_{\sqrt{\pi}}H_\ell\|.\end{aligned}$$

For $H=\Psi^*\Psi-\mathbb{I}_K$ we have the bound

$$\|D_{\sqrt{\pi}}H_\ell\| = \|D_{\sqrt{\pi}}(\Psi^*\psi_\ell - e_\ell)\| \le \|D_{\sqrt{\pi}}\Psi^*\psi_\ell\| \le \|D_{\sqrt{\pi}}\Psi^*\| = \|\Psi D_{\sqrt{\pi}}\|,$$

so setting $A=\Psi$ yields (e) while setting $A=(H_\ell)^*=(He_\ell)^*$ together with $\mathcal{G}=\{I:|I|=S\}$ yields (f). Finally, we can write $V_IV_I^*=VR_I^*R_IV^*=V[\operatorname{diag}(\mathbf{1}_{I\setminus\{\ell\}})+e_\ell e_\ell^*]V^*$, to get

$$\begin{aligned}\|\mathbb{E}[V_IV_I^*\cdot\mathbb{1}_I(\ell)]\| &\le \|VD_{\sqrt{\pi}}\mathbb{E}[D_{\sqrt{\pi}}^{-1}\operatorname{diag}(\mathbf{1}_{I\setminus\{\ell\}})D_{\sqrt{\pi}}^{-1}\cdot\mathbb{1}_I(\ell)]D_{\sqrt{\pi}}V^*\| + \|v_\ell v_\ell^*\mathbb{E}[\mathbb{1}_I(\ell)]\|\\ &\le \pi_\ell\cdot\|VD_{\sqrt{\pi}}\|^2+\pi_\ell\cdot\|v_\ell\|^2,\end{aligned}$$

which completes the proof of (g). **(d-g) ✓**

(h) We first prove that for a general matrix $\mathit{И}$ with zero diagonal we have

$$\begin{aligned}&\|\mathbb{E}[D_{\sqrt{\pi}}^{-1}\mathbb{I}_{\ell^c}R_I^*\mathit{И}_{I,I}\mathit{И}_{I,I}^*R_I\mathbb{I}_{\ell^c}D_{\sqrt{\pi}}^{-1}\cdot\mathbb{1}_I(\ell)]\|\\ &\quad\le\frac{\pi_\ell}{1-\pi_\ell}\left(3\|D_{\sqrt{\pi}}\mathit{И}e_\ell\|^2+\max_k\mathit{И}_{k\ell}^2+\tfrac{9}{2}\|D_{\sqrt{\pi}}\mathit{И}D_{\sqrt{\pi}}\|^2+\tfrac{3}{2}\max_k\|e_k^*\mathit{И}D_{\sqrt{\pi}}\|^2\right).\end{aligned}\tag{4.106}$$

Using (4.105) and the identity $\mathbb{I}_{\ell^c}=\operatorname{diag}(\mathbf{1}_{[K]\setminus\{\ell\}})$, we first rewrite

$$\begin{aligned}\mathbb{I}_{\ell^c}R_I^*\mathit{И}_{I,I}\mathit{И}_{I,I}^*R_I\mathbb{I}_{\ell^c}\cdot\mathbb{1}_I(\ell) &= (R_I^*\mathit{И}_{I,I}\mathit{И}_{I,I}^*R_I)\odot(\mathbf{1}_{[K]\setminus\{\ell\}}\mathbf{1}_{[K]\setminus\{\ell\}}^*)\cdot\mathbb{1}_I(\ell)\\ &=\sum_k(\mathit{И}_k\mathit{И}_k^*)\odot\big(\mathbf{1}_{I\setminus\{k,\ell\}}\mathbf{1}_{I\setminus\{k,\ell\}}^*\cdot\mathbb{1}_I(\ell)\mathbb{1}_I(k)\big).\end{aligned}$$

As before an application of the Schur Product Theorem and Theorem 4.17(e) leads to

$$\begin{aligned}\mathbb{E}_S[\mathbb{I}_{\ell^c}R_I^*\mathit{И}_{I,I}\mathit{И}_{I,I}^*R_I\mathbb{I}_{\ell^c}\cdot\mathbb{1}_I(\ell)] &= \sum_k(\mathit{И}_k\mathit{И}_k^*)\odot\mathbb{E}_S\big[\mathbf{1}_{I\setminus\{k,\ell\}}\mathbf{1}_{I\setminus\{k,\ell\}}^*\cdot\mathbb{1}_I(\ell)\mathbb{1}_I(k)\big]\\ &\preceq\frac{\pi_S(\ell)}{1-\pi_S(\ell)}(\mathit{И}_\ell\mathit{И}_\ell^*)\odot\mathbb{E}_{S-1}[\mathbf{1}_I\mathbf{1}_I^*]+\sum_{k\neq\ell}\frac{\pi_S(\ell)}{1-\pi_S(\ell)}\left(\mathit{И}_k\frac{\pi_S(k)}{1-\pi_S(k)}\mathit{И}_k^*\right)\odot\mathbb{E}_{S-2}[\mathbf{1}_I\mathbf{1}_I^*]\\ &\preceq\frac{\pi_S(\ell)}{1-\pi_S(\ell)}\Big((\mathit{И}_\ell\mathit{И}_\ell^*)\odot\mathbb{E}_{S-1}[\mathbf{1}_I\mathbf{1}_I^*]+M\odot\mathbb{E}_{S-2}[\mathbf{1}_I\mathbf{1}_I^*]\Big),\end{aligned}$$

where again $M=\mathit{И}\operatorname{diag}(\frac{\pi_S}{1-\pi_S})\mathit{И}^*$. Finally, Theorem 4.19, Theorem 4.17(b) and similar simplifications to above yield

$$\begin{aligned}&\|\mathbb{E}_S[D_{\sqrt{\pi}}^{-1}\mathbb{I}_{\ell^c}R_I^*\mathit{И}_{I,I}\mathit{И}_{I,I}^*R_I\mathbb{I}_{\ell^c}D_{\sqrt{\pi}}^{-1}\cdot\mathbb{1}_I(\ell)]\|\\ &\quad\le\frac{\pi_\ell}{1-\pi_\ell}\left(\|(D_{\sqrt{\pi}}^{-1}\mathit{И}_\ell\mathit{И}_\ell^*D_{\sqrt{\pi}}^{-1})\odot\mathbb{E}_{S-1}[\mathbf{1}_I\mathbf{1}_I^*]\|+\|(D_{\sqrt{\pi}}^{-1}MD_{\sqrt{\pi}}^{-1})\odot\mathbb{E}_{S-2}[\mathbf{1}_I\mathbf{1}_I^*]\|\right)\\ &\quad\le\frac{\pi_\ell}{1-\pi_\ell}\left(3\|D_{\sqrt{\pi}}\mathit{И}_\ell\mathit{И}_\ell^*D_{\sqrt{\pi}}\|+\|\operatorname{diag}(\mathit{И}_\ell\mathit{И}_\ell^*)\|+3\|D_{\sqrt{\pi}}MD_{\sqrt{\pi}}\|+\|\operatorname{diag}(M)\|\right)\\ &\quad\le\frac{\pi_\ell}{1-\pi_\ell}\left(3\|D_{\sqrt{\pi}}\mathit{И}e_\ell\|^2+\max_k\mathit{И}_{k\ell}^2+\tfrac{9}{2}\|D_{\sqrt{\pi}}\mathit{И}D_{\sqrt{\pi}}\|^2+\tfrac{3}{2}\max_k\|e_k^*\mathit{И}D_{\sqrt{\pi}}\|^2\right),\end{aligned}$$

which completes the proof of (4.106). To prove (h) note that for $\mathit{И} = H = \Psi^*\Psi - \mathbb{I}_K$ we have

$$\|D_{\sqrt{\pi}}He_k\| = \|e_k^*HD_{\sqrt{\pi}}\| = \|(\psi_k^*\Psi - e_k)D_{\sqrt{\pi}}\| \le \|\psi_k^*\Psi D_{\sqrt{\pi}}\| \le \|\Psi D_{\sqrt{\pi}}\|$$
$$\text{and} \qquad \|D_{\sqrt{\pi}}HD_{\sqrt{\pi}}\| \le \|D_{\sqrt{\pi}}\Psi^*\Psi D_{\sqrt{\pi}}\| \le \|\Psi D_{\sqrt{\pi}}\|^2$$

Since for $k \neq \ell$ we have $H_{k\ell}^2 = |\langle\psi_k, \psi_\ell\rangle|^2 \le \mu(\Psi)^2$ whenever $\|\Psi D_{\sqrt{\pi}}\| \le 1/3$ we have

$$\begin{aligned}\|\mathbb{E}[D_{\sqrt{\pi}}^{-1}\mathbb{I}_{\ell^c}R_I^*H_{I,I}H_{I,I}^*R_I\mathbb{I}_{\ell^c}D_{\sqrt{\pi}}^{-1}\cdot\mathbb{1}_I(\ell)]\| &\le \tfrac{3}{2}\cdot\pi_\ell\cdot\big(\mu(\Psi)^2 + \|\Psi D_{\sqrt{\pi}}\|^2\cdot\tfrac{9}{2}(1+\tfrac{1}{9})\big)\\ &\le 9\cdot\pi_\ell\cdot\max\{\mu(\Psi), \|\Psi D_{\sqrt{\pi}}\|\}^2,\end{aligned}$$

which completes the proof of (h). **(h) ✓**

(i-k) We first observe that using Theorem 4.7, we get the bound

$$\begin{aligned}&\|\mathbb{E}[\big(A(I)+B(I)\big)\big(A(I)+B(I)\big)^*]\|\\ &\quad\le \|\mathbb{E}[A(I)A(I)^*]\| + 2\|\mathbb{E}[A(I)\cdot\mathbb{I}\cdot B(I)^*]\| + \|\mathbb{E}[B(I)B(I)^*]\|\\ &\quad\le \|\mathbb{E}[A(I)A(I)^*]\| + 2\|\mathbb{E}[A(I)A(I)^*]\|^{\frac12}\|\mathbb{E}[B(I)B(I)^*]\|^{\frac12} + \|\mathbb{E}[B(I)B(I)^*]\|\\ &\quad= \Big(\|\mathbb{E}[(A(I)A(I)^*]\|^{\frac12} + \|\mathbb{E}[(B(I)B(I)^*]\|^{\frac12}\Big)^2. \qquad (4.107)\end{aligned}$$

Applying this to $\|\mathbb{E}[\Psi_I H_{I,I}H_{I,I}^*\Psi_I^*\cdot\mathbb{1}_I(\ell)]\|$ for the split

$$\Psi_I H_{I,I} = \Psi R_I^*H_{I,I} = \Psi(e_\ell e_\ell^* + I_{\ell^c})R_I^*H_{I,I} = \psi_\ell H_{\ell,I} + \Psi I_{\ell^c}R_I^*H_{I,I},$$

and using the symmetry of H, meaning $H_{\ell,I} = (H_{I,\ell})^*$, we get using (g) and (h) and the bounds $\mu(\Psi), \|\Psi D_{\sqrt{\pi}}\| \le 1/8$

$$\begin{aligned}&\|\mathbb{E}[{\Psi_I}^* H_{I,I}H_{I,I}^*\Psi_I^*\cdot\mathbb{1}_I(\ell)]\|^{\frac12}\\ &\quad\le \|\psi_\ell\mathbb{E}[(H_{I,\ell})^*H_{I,\ell}\cdot\mathbb{1}_I(\ell)]\psi_\ell^*\|^{\frac12} + \|\Psi\mathbb{E}[I_{\ell^c}R_I^*H_{I,I}H_{I,I}^*R_II_{\ell^c}\cdot\mathbb{1}_I(\ell)]\Psi^*\|^{\frac12}\\ &\quad\le \sqrt{\pi_\ell}\cdot\|\Psi D_{\sqrt{\pi}}\| + \|\Psi D_{\sqrt{\pi}}\|\cdot\|\mathbb{E}[D_{\sqrt{\pi}}^{-1}I_{\ell^c}R_I^*H_{I,I}H_{I,I}^*R_II_{\ell^c}D_{\sqrt{\pi}}^{-1}\cdot\mathbb{1}_I(\ell)]\|^{\frac12}\\ &\quad\le \sqrt{\pi_\ell}\cdot\|\Psi D_{\sqrt{\pi}}\| + \|\Psi D_{\sqrt{\pi}}\|\cdot 3\cdot\sqrt{\pi_\ell}\cdot\max\{\mu(\Psi), \|\Psi D_{\sqrt{\pi}}\|\}\\ &\quad\le \sqrt{2\pi_\ell}\cdot\|\Psi D_{\sqrt{\pi}}\|.\end{aligned}$$

To prove (j) we set $\mathcal{E} = \operatorname{diag}(\Psi^*Z)$, $\mathit{И} = \Psi^*Z - \operatorname{diag}(\Psi^*Z)$, and apply (4.107) to the split

$$D_{\sqrt{\pi}}^{-1}\mathbb{I}_{\ell^c}R_I^*\Psi_I^*Z_I = D_{\sqrt{\pi}}^{-1}\mathbb{I}_{\ell^c}R_I^*\mathcal{E}_{I,I} + D_{\sqrt{\pi}}^{-1}\mathbb{I}_{\ell^c}R_I^*\mathit{И}_{I,I}$$

which using (4.106) yields

$$\begin{aligned}&\|\mathbb{E}[D_{\sqrt{\pi}}^{-1}\mathbb{I}_{\ell^c}R_I^*\Psi_I^*Z_IZ_I^*\Psi_IR_I\mathbb{I}_{\ell^c}D_{\sqrt{\pi}}^{-1}\cdot\mathbb{1}_I(\ell)]\|^{\frac12}\\ &\quad\le \|\mathbb{E}[D_{\sqrt{\pi}}^{-1}\mathbb{I}_{\ell^c}R_I^*\mathcal{E}_{I,I}^2R_I\mathbb{I}_{\ell^c}D_{\sqrt{\pi}}^{-1}\cdot\mathbb{1}_I(\ell)]\|^{\frac12} + \|\mathbb{E}[D_{\sqrt{\pi}}^{-1}\mathbb{I}_{\ell^c}R_I^*\mathit{И}_{I,I}(\mathit{И}_{I,I})^*R_I\mathbb{I}_{\ell^c}D_{\sqrt{\pi}}^{-1}\cdot\mathbb{1}_I(\ell)]\|^{\frac12}\\ &\quad\le \sqrt{\pi_\ell}\cdot\|\mathcal{E}\| + \sqrt{\tfrac{\pi_\ell}{1-\pi_\ell}}\cdot\Big(3\|D_{\sqrt{\pi}}\mathit{И}e_\ell\|^2 + \max_k \mathit{И}_{k\ell}^2 + \tfrac{9}{2}\|D_{\sqrt{\pi}}\mathit{И}D_{\sqrt{\pi}}\|^2 + \tfrac{3}{2}\max_k\|e_k^*\mathit{И}D_{\sqrt{\pi}}\|^2\Big)^{\frac12}\\ &\quad\le \sqrt{\pi_\ell}\cdot\tfrac{\varepsilon^2}{2} + \sqrt{\tfrac{\pi_\ell}{1-\pi_\ell}}\cdot\Big(3\|D_{\sqrt{\pi}}\Psi^*z_\ell\|^2 + \max_{k\neq\ell}|\langle\psi_k, z_\ell\rangle|^2\\ &\qquad\qquad + \tfrac{9}{2}(\|D_{\sqrt{\pi}}\Psi^*ZD_{\sqrt{\pi}}\| + \|D_{\sqrt{\pi}}\mathcal{E}D_{\sqrt{\pi}}\|)^2 + \tfrac{3}{2}\max_k\|\psi_k^*ZD_{\sqrt{\pi}}\|^2\Big)^{\frac12}\\ &\quad\le \sqrt{\pi_\ell}\cdot\tfrac{\varepsilon^2}{2} + \sqrt{\tfrac{\pi_\ell}{1-\pi_\ell}}\cdot\Big(3\|\Psi D_{\sqrt{\pi}}\|^2\varepsilon^2 + \varepsilon^2 + \tfrac{9}{2}(\|\Psi D_{\sqrt{\pi}}\|\|ZD_{\sqrt{\pi}}\| + \|\pi\|_\infty\tfrac{\varepsilon^2}{2})^2 + \tfrac{3}{2}\|ZD_{\sqrt{\pi}}\|^2\Big)^{\frac12}.\end{aligned}$$

Inserting the bounds $\varepsilon \leq \sqrt{2}$, $\|\Psi D_{\sqrt{\pi}}\| \leq 1/8$ and $\|\pi\|_\infty \leq 1/3$ we finally arrive at

$$\begin{aligned}
&\|\mathbb{E}[D_{\sqrt{\pi}}^{-1}\mathbb{I}_{\ell^c} R_I^* \Psi_I^* Z_I Z_I^* \Psi_I R_I \mathbb{I}_{\ell^c} D_{\sqrt{\pi}}^{-1} \cdot \mathbb{1}_I(\ell)]\|^{\frac{1}{2}} \\
&\quad \leq \sqrt{\pi_\ell} \cdot \max\{\varepsilon, \|ZD_{\sqrt{\pi}}\|\} \cdot \left[\tfrac{1}{\sqrt{2}} + \sqrt{\tfrac{3}{2}} \cdot \left(\tfrac{3}{8^2} + 1 + \tfrac{9}{2}(\tfrac{1}{8} + \tfrac{1}{3\sqrt{2}})^2 + \tfrac{3}{2}\right)^{\frac{1}{2}}\right] \\
&\quad \leq \sqrt{\pi_\ell} \cdot \max\{\varepsilon, \|ZD_{\sqrt{\pi}}\|\} \cdot 3
\end{aligned}$$

Reversing the roles of Z and Ψ in the inequalities above and simplifications using the same bounds $\varepsilon, \|\Psi D_{\sqrt{\pi}}\|, \|\pi\|_\infty$ straightforwardly leads to (k). **(i-k)✓**

■

4.10. Discussion

We have shown that both aK-SVD and MOD converge to the generating dictionary under very general conditions and a non-uniform sparse supports signal model. Even though these algorithms arise from different approaches to solving the minimisation problem that is dictionary learning, they surprisingly share the same structure in the dictionary update step. We also suspect that the ideas of this proof could be reused to show convergence of the ITKrM algorithm as well.

Chapter 5

Average performance of OMP and Thresholding under dictionary mismatch

The following chapter essentially is a reprint of the article

M. Pali, S. Ruetz, and K. Schnass. Average performance of OMP and Thresholding under dictionary mismatch. IEEE Signal Processing Letters, 29:1077–1081, 2022

https://doi.org/10.1109/LSP.2022.3167313

In the previous chapter we derived sufficient conditions that the dictionary learning algorithms MOD and A-K-SVD recover a generating dictionary with high probability. The keen reader might have noticed that we only looked at Thresholding for the sparse approximation step and might ask him or herself why OMP was not included in the analysis. This nevertheless is not without good reason, and in this chapter we provide the theoretical and numerical justification for this choice.

5.1. Introduction

For convenience we recall the ideas of dictionary learning, OMP and Thresholding in order to make the results more accessible. Recall that in dictionary learning the goal is to decompose a data matrix $Y = (y_1, \dots, y_N) \in \mathbb{R}^{d \times N}$ into a dictionary matrix $\Phi = (\phi_1, \dots, \phi_K) \in \mathbb{R}^{d \times K}$, where each column (also called atom) is normalised, i.e., $\|\phi_k\|_2 = 1$, and a sparse coefficient matrix $X = (x_1, \dots, x_N) \in \mathbb{R}^{K \times N}$ such that $Y \approx \Phi X$ and X is column-wise sparse. This problem can be formulated as an optimisation program

$$\min_{\Phi \in \mathcal{D}_K} \sum_{n=1}^{N} \min_{\|x_n\|_0 \leq S} \|y_n - \Phi x_n\|_2^2, \tag{5.1}$$

where $\mathcal{D}_K$ denotes the set of all dictionaries with K normalised atoms and $\|x\|_0$ counts the number of non-zero entries of a vector x. While sparse representations are important for performing many signal processing tasks such as denoising [32] or data reconstruction from incomplete information [51], solving a highly non-convex minimisation problem as in (5.1) is notoriously difficult [79]. Some of the most used dictionary learning algorithms belong to the class of alternating optimisation algorithms, which alternate between updating the sparse coefficient matrix X while fixing the dictionary Φ and updating the dictionary Φ while fixing X. Popular examples include K-SVD [6], MOD [33], ITKrM [72] or the neural algorithms in [8]. Even updating the sparse coefficient matrix X, meaning finding the best sparse approximation of each signal y_n in Φ, is generally NP-hard unless the dictionary forms an orthonormal system [37, 53]. In particular, in sparse approximation we want to approximate a given signal $y \in \mathbb{R}^d$ by a linear combination of only a small number $S \ll d$ of elements $\phi_i \in \mathbb{R}^d$ out of some given dictionary $\Phi = (\phi_1, \dots, \phi_K)$. This means, denoting the restriction of Φ and x to the columns resp. entries indexed by some set I by Φ_I and x_I, we want to find

$$y \approx \sum_{k \in I} \phi_k x_k = \Phi_I x_I \quad \text{such that} \quad |I| = S \ll d. \tag{5.2}$$

The problem of finding the best S-sparse approximation of y in Φ, meaning the best S-support I and coefficient vector x, however, is combinatorial. To approximate its solution efficiently, suboptimal routines that avoid searching through all possible sets I are typically used. One of the most practically used sparse approximation algorithms is Orthogonal Matching Pursuit (OMP) [59]. OMP finds the support iteratively by adding the index of the atom which has the largest absolute inner product with the residual and updating the residual. In particular, initialising with $r_J = y$ and $J = \emptyset$, it

$$\begin{aligned} \text{finds} \quad & i \in \arg\max_k |\langle \phi_k, r_J \rangle| \quad \text{and} \\ \text{updates} \quad & J \leftarrow J \cup \{i\} \quad \text{resp.} \quad r_J = y - P(\Phi_J) y, \end{aligned}$$

where $P(\Phi_J)$ denotes the projection onto the span of atoms indexed by J, iterating until a stopping criterion is met. As the projection can be calculated iteratively the computational cost is determined by the K inner products $\langle \phi_k, r_J \rangle$ in each iteration, combining to an overall cost of $\mathcal{O}(SdK)$ [64].

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

$$I \in \arg\max_{J:|J|=S} \|\Phi_J^\top y\|_1, \tag{5.3}$$

meaning it simply chooses those atoms which yield the S largest inner products in absolute value with the signal. Together with the projection this combines to a much reduced computational complexity of $\mathcal{O}(dK + S^3)$.
Over the years, quite a few results about sufficient conditions for OMP and Thresholding to recover the correct support emerged [80, 74]. Most recently, in [73] average case results for OMP were derived. However, these results assume exact knowledge of the generating dictionary, whereas in practice only an approximation might be available. Hence, they do not apply directly. In dictionary learning, for example, the initial guess (and also the subsequent updates) are naturally quite different from the signal generating dictionary. Thus results for Thresholding under dictionary mismatch can be found implicitly in several dictionary learning papers [8, 72, 56]. Further, a similar problem known as basis mismatch is studied in compressed

sensing [24, 11]. There the dictionary rather than the signal coefficients is usually assumed to be random.

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

5.2. Setting

Definition 5.1 (Signal model) *Given a $d \times K$ dictionary Φ, we assume that our noisy signals are generated as*

$$y = \Phi_I x_I + \eta = \sum_{i \in I} \phi_i \sigma_i c_{p(i)} + \eta, \tag{5.4}$$

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

This signal model is quite general. Using Rademacher signs σ_i simply ensures that the coefficients x_i are centered, which together with boundedness is a quite common assumption [8, 21]. Further, we want to point out that sub-Gaussian noise includes both bounded and Gaussian noise. Choosing I uniformly at random among all sets of size S allows us to conclude that for any dictionary Ψ with small operator norm $\|\Psi\|_{2,2}$ and small coherence $\mu(\Psi)$ we have $\vartheta(\Psi_I) \leq 1/2$ with high probability [25, 83]. This could be replaced by a more general non-uniform sampling scheme, where similar conditions on Ψ including suitable weights again lead to $\vartheta(\Phi_I) \leq 1/2$ with high probability - see Chapter 2.

We recall some notation from Chapter 1, which will be used throughout this chapter. For a perturbed version Ψ of a generating dictionary Φ, we set $Z := \Phi - \Psi$ and define its distance to Φ as $\varepsilon := \max_i \|z_i\|_2$. The perturbation parameter $\nu := \max_{i,j} |\langle \psi_i, z_j \rangle|$ measures how correlated the perturbation of one atom is with the other perturbed atoms. Finally, for a vector $v \in \mathbb{R}^K$ and an index ℓ, we define $v_{\geq \ell} := v_I$ for $I = \{\ell, \dots, K\}$.

5.3. Main results

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

Theorem 5.2 *Assume the signals are generated following the model in (5.4) with signal generating dictionary Φ and let Ψ be a perturbed version of Φ with parameter ν.*
***OMP:** Let $\ell \leq S$. If Ψ satisfies*

$$\mu(\Psi) \leq \frac{1}{4n \log K} \quad \text{and} \quad \|\Psi\|_{2,2}^2 \leq \frac{K}{16ne^2 S \log K}, \tag{5.5}$$

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

$$\frac{1-\gamma}{2} > \mu(\Psi)\left(\max_{i\le\ell}\frac{\|c_{\ge i}\|_1}{c_i} + \sqrt{\ell}\max_{i\le\ell}\cdot\frac{\|c_{\ge i}\|_2}{c_i}\right) + (1+2\ell\mu(\Psi))\sqrt{2n\log K}\left(\frac{\nu\|c\|_2+\rho}{c_\ell}\right), \tag{5.6}$$

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

$$\frac{1-\gamma}{2} \ge \left(\mu(\Psi)\cdot\frac{\|c\|_2}{c_\ell} + \frac{\nu\|c\|_2+\rho}{c_\ell}\right)\sqrt{2n\log K}, \tag{5.7}$$

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

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

$$|\langle \psi_\ell, r_J\rangle| > \max_{i\in R}|\langle \psi_i, r_J\rangle|. \tag{5.8}$$

Rewriting $y = \Phi_I x_I + \eta = \Psi_I x_I + Z_I x_I + \eta$, and abbreviating $Q(\Psi_J) = \mathbb{I} - P(\Psi_J)$ we get

$$r_J = Q(\Psi_J)\Psi_L x_L + Q(\Psi_J)Z_I x_I + Q(\Psi_J)\eta,$$

so in order to bound $|\langle \psi_\ell, r_J\rangle|$ from below, we need to bound the inner products of ψ_ℓ with the terms on the r.h.s above. First note that by [25, Theorem 3.1] and (5.5) $\vartheta(\Psi_J) \le \vartheta(\Psi_I) \le 1/2$ except with probability $216K^{1-n}$. So for $\bar{L} = L \setminus \{\ell\}$ we have

$$\begin{aligned}
&|\langle \psi_\ell, \Psi_L x_L - P(\Psi_J)\Psi_L x_L\rangle| \\
&\ge |x_\ell| - \left|\langle \Psi_{\bar{L}}^\top \psi_\ell, x_{\bar{L}}\rangle\right| - \left|\langle \Psi_J^\top \psi_\ell, (\Psi_J^\top\Psi_J)^{-1}\Psi_J^\top\Psi_L x_L\rangle\right| \\
&\ge \|x_L\|_\infty - \|\Psi_{\bar{L}}^\top\|_\infty\|x_{\bar{L}}\|_1 \\
&\qquad - \|\Psi_J^\top\psi_\ell\|_2\|(\Psi_J^\top\Psi_J)^{-1}\|_{2,2}\|\Psi_J^\top\Psi_L\|_{2,2}\|x_L\|_2 \\
&\ge \|x_L\|_\infty - \mu(\Psi)\|x_L\|_1 - \mu(\Psi)\sqrt{|J|}\frac{\vartheta(\Psi_I)}{1-\vartheta(\Psi_I)}\|x_L\|_2.
\end{aligned}$$

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

$$\begin{aligned}
&|\langle \psi_\ell, \Psi_L x_L - P(\Psi_J)\Psi_L x_L\rangle| \\
&\qquad \le \gamma\|x_L\|_\infty + \mu(\Psi)\|x_L\|_1 + \mu(\Psi)\sqrt{|J|}\|x_L\|_2.
\end{aligned}$$

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

$$\begin{aligned}
&\left|\langle \psi_i, Z_I x_I - \Psi_J(\Psi_J^\top\Psi_J)^{-1}\Psi_J^\top Z_I x_I\rangle\right| \\
&\le |\langle \psi_i, Z_I x_I\rangle| + \frac{\|\Psi_J^\top\psi_i\|_2}{1-\vartheta(\Psi_J)}\cdot\sqrt{|J|}\cdot\|\Psi_J^\top Z_I x_I\|_\infty \\
&\le \max_i|\langle \psi_i, Z_I x_I\rangle|\cdot(1+2|J|\mu(\Psi)).
\end{aligned} \tag{5.9}$$

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

$$\begin{aligned}\mathbb{P}(|\langle\psi_i, Z_I x_I\rangle| > t) &= \mathbb{P}\big(|\textstyle\sum_j \langle\psi_i, z_j\rangle c_{p(j)}\sigma_j| > t\big) \\ &\leq 2\exp\left(\frac{-t^2}{2\sum_j \langle\psi_i, z_j\rangle^2 x_j^2}\right) \leq 2\exp\left(\frac{-t^2}{2\nu^2\|x_I\|_2^2}\right).\end{aligned}$$

Setting $t = t_\nu := \nu\|x_I\|_2\sqrt{2n\log K}$ we get via a union bound that $\max_i |\langle\psi_i, Z_I x_I\rangle| < t_\nu$ except with probability $2K^{1-n}$.
Simply replacing $Z_I x_I$ by η in (5.9) we further get

$$|\langle\psi_i, Q(\Psi_J)\eta\rangle| \leq \max_i |\langle\psi_i, \eta\rangle| \cdot (1 + 2|J|\mu(\Psi)).$$

Since η is sub-Gaussian, Markov's inequality leads to $\mathbb{P}(|\langle\psi_i, \eta\rangle| > t) \leq 2e^{-t^2/(2\rho^2)}$. Setting $t = t_\rho := \rho\sqrt{2n\log K}$ and a union bound yield that $\max_i |\langle\psi_i, \eta\rangle| \leq t_\rho$ except with probability $2K^{1-n}$.
After collecting all our bounds into (5.8) and rearranging, we get the following sufficient condition for OMP to pick another correct atom except with probability $(216 + 2 + 2) \cdot K^{1-n}$

$$\begin{aligned}\frac{1-\gamma}{2} > \mu(\Psi)\left(\frac{\|x_L\|_1}{\|x_L\|_\infty} + \sqrt{|J|}\frac{\|x_L\|_2}{\|x_L\|_\infty}\right) \\ + \left(\frac{\nu\|x_I\|_2 + \rho}{\|x_L\|_\infty}\right)(1 + 2|J|\mu(\Psi))\sqrt{2n\log K}.\end{aligned}$$

To get the final result observe that $\|x_I\|_2 = \|c\|_2$ and that in the ℓ-th step $|J| = \ell - 1$ and $\|x_L\|_\infty \geq c_\ell$. If $\|x_L\|_\infty = c_i$ for the smallest possible i, then $\|x_L\|_p \leq \|c_{\geq i}\|_p$ and

$$\frac{\|x_L\|_p}{\|x_L\|_\infty} \leq \max_{i \leq \ell} \frac{\|c_{\geq i}\|_p}{c_i}.$$

To get the statement for thresholding, observe that

$$\langle\psi_i, y\rangle = x_i + \langle\psi_i, \Psi_{I\setminus\{i\}} x_{I\setminus\{i\}}\rangle + \langle\psi_i, Z_I x_I\rangle + \langle\psi_i, \eta\rangle.$$

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

$$|\langle\psi_i, y\rangle| \leq |x_i| + (\mu(\Psi)\|x\|_2 + \nu\|x\|_2 + \rho)\sqrt{2n\log K},$$

for all i as well as the corresponding lower bound, so (5.7) ensures that the inner products of atoms having coefficients $c_i \geq c_\ell$ are larger than those having coefficients $c_i \leq \gamma c_\ell$. ■

5.4. Comparison of OMP and Thresholding

In the perturbation and noise-free case our result reduces to that from [73], showing that the recovery condition for OMP becomes easier to fulfill if we have decaying coefficients. So for constant coefficients, we need $\mu(\Psi)S \lesssim 1$, while for coefficients forming a geometric sequence, meaning $c_i = \alpha^i$ for $\alpha \in (0, 1)$ and $i \leq S$, we only need $\mu(\Psi) \lesssim 1 - \alpha$ as well as $\mu^2(\Psi)S \lesssim 1 - \alpha^2$ for full recovery. Unfortunately, in the case of perturbations, as ν grows, this advantage turns

into a disadvantage, since the term $\|c\|_2/c_S$ grows with faster decay, e.g. equaling $\sqrt{S}$ for constant coefficients and $\alpha^{-S}/\sqrt{1-\alpha^2}$ for the geometric sequence.
For thresholding on the other hand the term scaled by ν, which grows with coefficient decay, already appears in the perturbation- and noise-free recovery-condition. This means that thresholding never performs well with large coefficient decay but also that its performance does not degrade dramatically with perturbations.

To better judge the influence of the perturbation parameter ν, we have a look at its extreme and typical size. For reasonable perturbation sizes, $\varepsilon := \max_k \|z_k\|_2 \leq 0.7$, we have $\nu = \max_{i,j} |\langle \psi_i, z_j \rangle| \approx \max_{i \neq j} |\langle \phi_i, z_j \rangle|$, so at worst, if $z_j \approx \varepsilon \phi_k$, we have $\nu \approx \varepsilon$. On the other hand for random (rescaled Gaussian) perturbations we have $\nu \approx \varepsilon \sqrt{\log K/d}$. Also a more involved analysis – beyond the scope of this chapter – for uniformly distributed supports, leads to a result corresponding to the above with $\nu \approx \|Z\|_{2,2}/\sqrt{K}$ [58].
To see how accurate our conditions are, we next conduct some numerical simulations in $\mathbb{R}^d$, for $d = 128$, for the case of geometric coefficient sequences and random perturbations. We assume that the signals follow the model in (5.4), where the support I is chosen uniformly at random. For $\alpha \in [0.75, 1]$ and $S \in \{2, \ldots, 54\}$ we set $c_i = \alpha^i$ for $i \leq S$ and $c_i = 0$ for all $i > S$. As generating dictionary Φ we use the concatenation of the Dirac and DCT bases. We obtain a perturbed dictionary Ψ with distance ε to Φ by setting $\psi_k = \left(1 - \varepsilon^2/2\right) \phi_k + \left(\varepsilon^2 - \varepsilon^4/4\right)^{1/2} v_k$, where v_k is drawn uniformly at random from the unit sphere orthogonal to ϕ_k. For our experiments we use $N = 1000$ signals per sparsity level and decay parameter. The results in Figure 5.1 show that OMP outperforms Thresholding — but only for very small perturbations. This performance gap closes with growing levels of perturbation. In order to compare the sufficient conditions in Theorem 5.2 to our empirical results we plot the following boundaries

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

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

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

for $\mu = \frac{1}{8} = \mu(\Phi) \approx \mu(\Psi)$ and $\nu = \varepsilon/\sqrt{d}$. These results confirm the behaviour discussed above and show that the conditions in Theorem 5.2 are rather tight (up to constants).

5.5. Dictionary learning using OMP and Thresholding

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

As can be seen in Figure 5.2, all combinations of algorithms were able to fully recover the dictionary. Interestingly, the increased complexity of OMP does not seem to provide an advantage over Thresholding in the first few iterations. In the noiseless case OMP starts

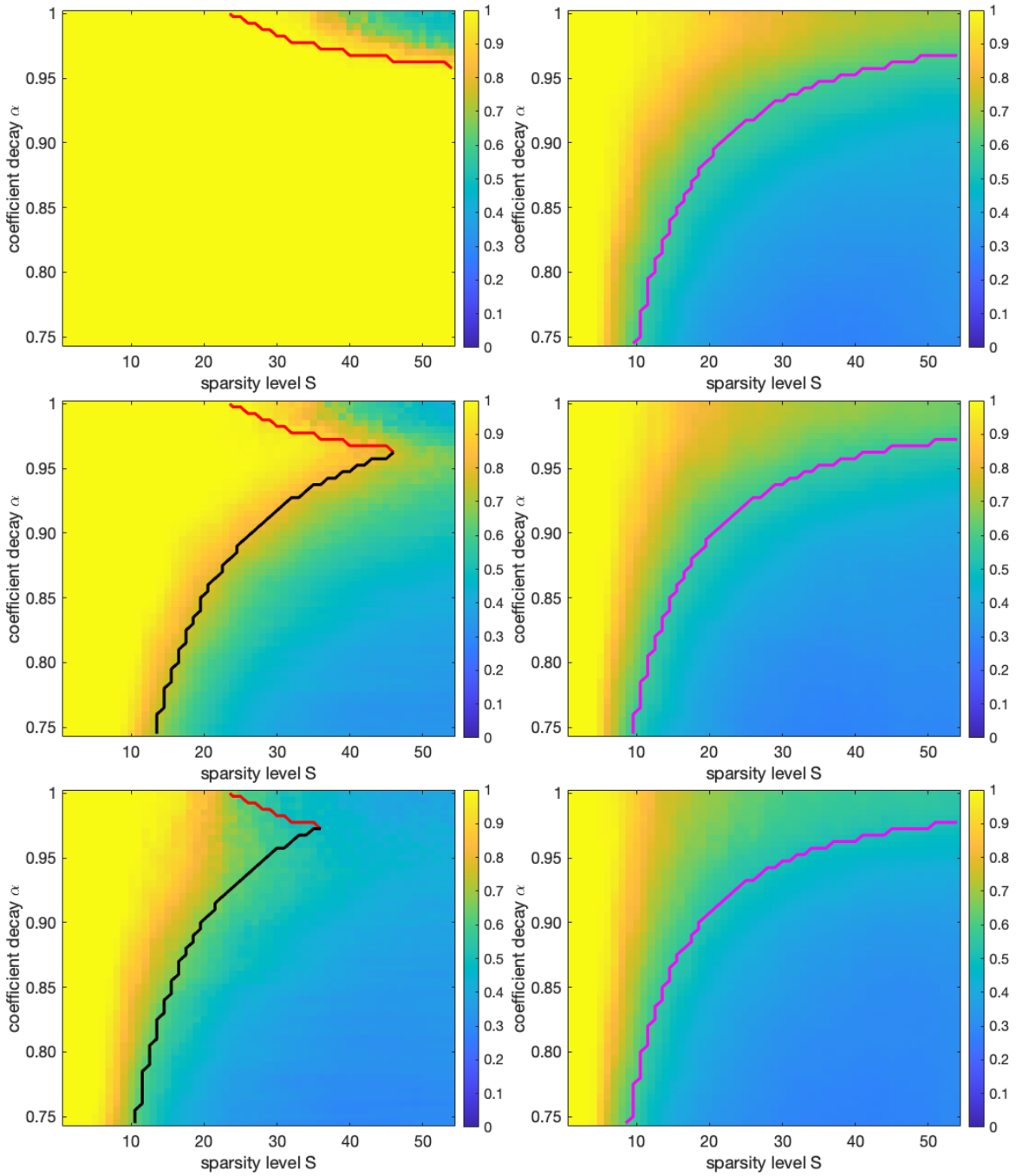

Figure 5.1: Average percentage of correctly recovered atoms via OMP (left column) and Thresholding (right column) with perturbed dictionary Ψ where $\varepsilon = 0$ (top), $\varepsilon = 0.2$ (middle) and $\varepsilon = 0.5$ (bottom), for noiseless signals with generating dictionary Φ, various sparsity levels and coefficient decay parameters. The red, black and magenta lines indicate the theoretical decision boundaries.

to outperform Thresholding only once the learned dictionary atoms are very close to their corresponding atoms in the generating dictionary, while in the noisy case, they perform nearly on par with each other. Taking into account that the Thresholding is computationally far less

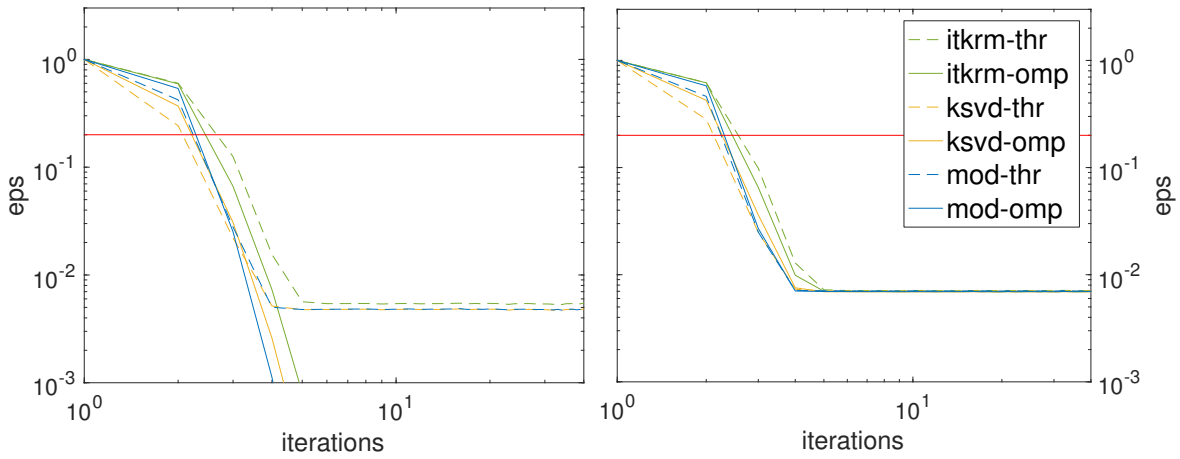

Figure 5.2: Average distance of atoms to the generating dictionary for various dictionary learning algorithms using OMP (full lines) and Thresholding (dashed lines) for a well-behaved initialisation with $\varepsilon = 1$ (Section 5.4), using noiseless (left) resp. noisy training signals (right). The red line indicates the error at which the inner products between learned atoms and generating atoms would equal 0.98.

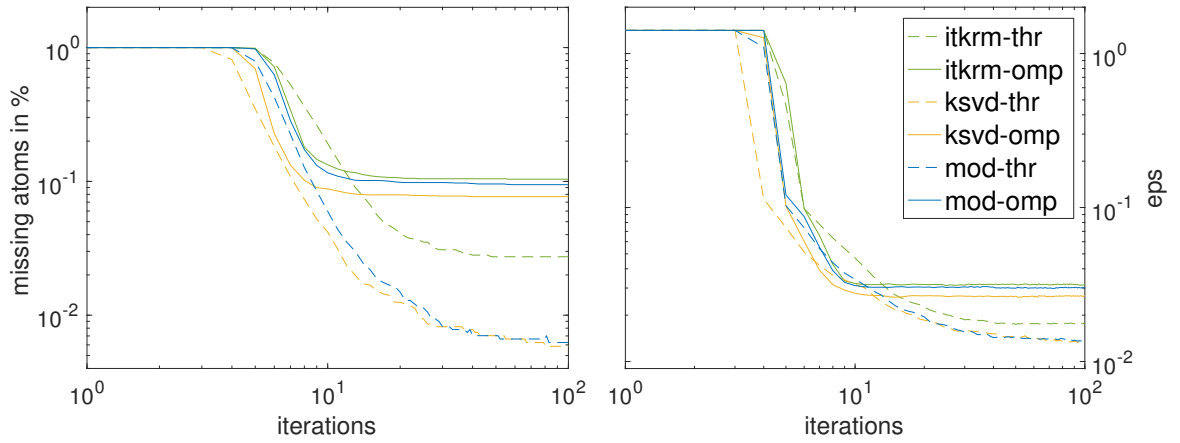

Figure 5.3: (left) Percentage of atoms not found and (right) average distance to the generating atoms on *found* atoms, both using OMP (full lines) and Thresholding (dashed lines).

demanding than OMP there might not be a benefit in employing OMP in dictionary learning — in the early stages.

Obviously an initialisation as defined in Section 5.4 is quite unrealistic, which is why we repeat the same experiment using the noisy signals with a fully random initialisation. The results in Figure 5.3 paint a far more accurate picture of reality. It can be seen that, contrary to the previous experiment, OMP is not able to find all atoms of the generating dictionary (left), whereas Thresholding is able to find almost all atoms. Note that we used the convention that ϕ_i is *found* if for a recovered ψ_k we have $|\langle \phi_i, \psi_k \rangle| \geq 0.99$, however the plot looks the same using 0.90 instead. Moreover, looking at the average distance of *found* atoms, we see that OMP is not able to outperform Thresholding either (right).

5.6. Discussion

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

Chapter 6

Discussion and Outlook

Now it is time to wrap things up. In Chapter 2 we have derived concentration inequalities for the operator norms of non-uniformly selected random submatrices. This has allowed us to derive sufficient conditions for sparse approximation algorithms to recover the sparse coefficients in a very general signal model. Building upon these results, in Chapter 3 we have derived optimal subsampling strategies in a compressed sensing setup. These subsampling strategies depend on the distribution of sparse supports, which can be estimated from data, yielding state of the art performance in numerical experiments. In Chapter 4, sufficient conditions for convergence of two popular dictionary learning algorithms were derived. It was shown that convergence happens under much more relaxed conditions as previous results suggested. In Chapter 5 we have argued that in settings where one only has access to a perturbed version of a dictionary, Thresholding is a viable alternative to OMP and that in some settings, like dictionary learning, Thresholding might even be preferable.
Together with the answers it provides, this thesis also opens up many fresh and challenging research directions. In Chapter 3 it was pointed out that it should be possible to weaken the assumptions on the random signs of the signal model by employing the so called golfing-scheme. This would generalise the result and maybe lead to improved constants in the conditions.
The most interesting questions regarding Chapter 3 come to mind when analysing the intricate relationship between the sampling density π and the probabilities p. Deriving lower bounds on the number of measurements in a blocks of measurements setting for a fixed underlying distribution p of the sparse supports would be very interesting and of practical relevance. As in the Fourier-Haar cases with measurements along vertical lines, other special cases of sensing matrices and block designs could be analysed together with different assumptions on the distribution p.
Recall that one of the main results of this chapter reads as $\pi_k = \frac{\max\{a_k D_\omega a_k^*, \|a_k\|_\infty^2\}}{L}$. Forgetting about the second term in the maximum, we see that we can write the vector π as a matrix vector product. Defining the matrix $\bar{A}$ as the entry-wise conjugate of A we have

$$\pi = (A \odot \bar{A})p.$$

In Chapter 3 we usually estimated p from data and derived π via the given formulas. But one can also turn this argument on its head and try to guess the underlying distribution p from a subsampling strategy π, since in practice it is very common to use some heuristically inspired subsampling density π. By solving the above system of linear equations (recall that the matrices A and $\bar{A}$ are known) one can get information about the implicit assumptions on the prior distribution p for any given subsampling density π. This information can then

be checked against the data and decisions about the applicability of a heuristically inspired subsampling strategy π can be made.
Other possible research directions emerged during the study of the dictionary learning algorithms in Chapter 4. We looked at two dictionary learning algorithms and essentially proved convergence of both of them with the help of the same technical lemmas and concentration inequalities. We therefore strongly suspect that results could easily be extended to other dictionary learning algorithms like the ITKrM or gradient descent schemes.
Another big step for theoretical dictionary learning would be partial support recovery. What do we mean by that? In general the true sparsity level of a certain signal or signal class is unknown and what we see in simulations is that both MOD and K-SVD do not rely too much on S. So a big step forward would be to show that even if the sparsity level is estimated smaller than the ground truth, the algorithms still recover the generating dictionary — or recover it at least partially.

List of Figures

2.1 Left: Original image from which the patches are extracted. Middle: Relative frequency of wavelet coefficients above threshold (blue) – average frequency (red) on a log scale. Right: Locations of non-zeros coefficients in the 2D Haar-Wavelet basis – the higher the row or column index the smaller the corresponding wavelet. 16

2.2 From left to right: The K-space $\{(k_1, k_2) : -\sqrt{K}/2 + 1 \leq k_1, k_2 \leq \sqrt{K}/2\}$ with the frequencies used for the measurement matrix A_1. The frequencies used for the measurement matrix A_2. Locations of non-zero coefficients of patches in the 2D-Haar Wavelet Basis. Expectation of each atom to be in the support (blue) and average expectations for comparison (red) on a log scale. 26

2.3 Left: Expectations of the Bernoulli random variables employed in our distribution models. Right: The same plot with the relative frequency of the wavelet coefficients from 2.1 for comparison. 28

2.4 Percentage of recovered supports (y-axis) for Thresholding with different sensing dictionaries for various sizes of sparse supports (x-axis). Blue corresponds to no sensing dictionary, red to the uniform average case sensing dictionary and orange to the distribution specific average case sensing dictionary. 29

2.5 Percentage of recovered supports (y-axis) for OMP with different sensing dictionaries for various sizes of sparse supports (x-axis). Blue corresponds to no sensing dictionary, red to the uniform average case sensing dictionary and orange to the distribution specific average case sensing dictionary. 30

2.6 Percentage of recovered supports (y-axis) for BP with different preconditioning strategies for various sizes of sparse supports (x-axis). Blue corresponds to the original ℓ_1-minimisation problem, red to preconditioning with uniform weights and orange to preconditioning with the correct weights. 30

3.1 Subsampling densities (top row) and corresponding samples (bottom row) for the adapted variable density sampling scheme (left column), the uniform distribution (middle right) and the coherence based subsampling scheme (right row). The resulting average PSNR are: Adapted - 133.5, Uniform - 105.6 and Coherence - 62.3. 42

3.2 Adapted variable density sampling scheme (left column) vs polynomial decay (middle column). Matrix W of sparse support distribution in the DB4 wavelet basis (top right) and test image (bottom right). The resulting PSNR values are: Adapted - 32.8 and Polynomial - 32.0. 43

3.3 Adapted variable density sampling scheme (left column) vs polynomial decay (middle column). Matrix W of sparse support distribution in the DB4 wavelet basis (top right) and test image (bottom right). The resulting PSNR values are: Adapted - 27.9 and Polynomial - 26.8. 43

3.4 Adapted variable density sampling scheme (left column) vs polynomial decay (middle column). Matrix W of sparse support distribution in the DB4 wavelet basis (top right) and test image (bottom right). The resulting PSNR values are: Adapted - 22.9 and Polynomial - 11.6. 44

3.5 Adapted variable density sampling schemes with vertical lines (left column) and squares (middle column). Matrix W of sparse support distribution in the separable 2D DB4 wavelet basis (top right), test image (bottom right) and reconstructions (bottom left and middle). The resulting PSNR values are: Lines - 29.9 and Squares - 33.9. 48

5.1 Comparison of correctly recovered atoms via OMP and Thresholding 103

5.2 Comparison of average distance of atoms via OMP and Thresholding 104

5.3 Comparison of average distance of atoms via OMP and Thresholding 104

List of Tables

2.1 Reconstruction error for two different sensing matrices 27

Bibliography

[1] B. Adcock, A.C. Hansen, and B. Roman. A note on compressed sensing of structured sparse wavelet coefficients from subsampled fourier measurements. *IEEE Signal Processing Letters*, 23(5):732–736, 2016.

[2] B. Adcock, A.C. Hansen, C. Poon, and B. Roman. Breaking the coherence barrier: A new theory for compressed sensing. *Forum of Mathematics, Sigma*, 5:e4, 2017. doi: 10.1017/fms.2016.32.

[3] B. Adcock, C. Boyer, and S. Brugiapaglia. On oracle-type local recovery guarantees in compressed sensing. *Information and Inference: A Journal of the IMA*, 10(1):1–49, 2020.

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

[5] A. Agarwal, A. Anandkumar, P. Jain, and P. Netrapalli. Learning sparsely used overcomplete dictionaries via alternating minimization. *SIAM Journal on Optimization*, 26 (4):2775–2799, 2016.

[6] 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.

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

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

[9] 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.

[10] S. Becker, J. Bobin, and E. Candès. Nesta: A fast and accurate first-order method for sparse recovery. *SIAM Journal on Imaging Sciences*, 4(1):1–39, 2011.

[11] Stéphanie Bernhardt, Rémy Boyer, Sylvie Marcos, and Pascal Larzabal. Compressed sensing with basis mismatch: Performance bounds and sparse-based estimator. *IEEE Transactions on Signal Processing*, 64(13):3483–3494, 2016.

[12] J. Bourgain and L. Tzafriri. Invertibility of "large" submatrices with applications to the geometry of banach spaces and harmonic analysis. *Israel J. Math*, 57(2):137–224, 1987.

[13] C. Boyer, P. Weiss, and J. Bigot. An algorithm for variable density sampling with block-constrained acquisition. *SIAM Journal on Imaging Sciences [electronic only]*, 7, 10 2013.

[14] C. Boyer, P. Weiss, and J. Bigot. Compressed sensing with structured sparsity and structured acquisition. *Applied and Computational Harmonic Analysis*, 46(2):312 – 350, 2019.

[15] M. Buda. Brain MRI segmentation. `https://www.kaggle.com/datasets/mateuszbuda/lgg-mri-segmentation`, 2019. Accessed: 2022-06-27.

[16] E. Candès and Y. Plan. Near-ideal model selection by l1 minimization. *The Annals of Statistics*, 37:2145–2177, 2009.

[17] 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.

[18] 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.

[19] E. J. Candes and Y. Plan. A probabilistic and ripless theory of compressed sensing. *IEEE Transactions on Information Theory*, 57(11):7235–7254, 2011.

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

[21] Niladri Chatterji and Peter L Bartlett. Alternating minimization for dictionary learning with random initialization. *Advances in Neural Information Processing Systems*, 30:1997–2006, 2017.

[22] N. Chauffert, P. Ciuciu, J. Kahn, and P. Weiss. Variable density sampling with continuous trajectories. *SIAM Journal on Imaging Sciences*, 7, 11 2013.

[23] N. Chauffert, P. Ciuciu, and P. Weiss. Variable density compressed sensing in MRI. Theoretical vs heuristic sampling strategies. In *2013 IEEE 10th International Symposium on Biomedical Imaging*, pages 298–301, 2013.

[24] Y. Chi, L. Scharf, A. Pezeshki, and R. Calderbank. Sensitivity to basis mismatch in compressed sensing. *IEEE Transactions on Signal Processing*, 59(5):2182–2195, 2011.

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

[26] I.Y. Chun and B. Adcock. Compressed sensing and parallel acquisition. *IEEE Transactions on Information Theory*, 63(8):4860–4882, 2017.

[27] W. Dai and Y. Ye. A characterization on singular value inequalities of matrices. *Journal of Function Spaces*, 2020:1–4, 2020.

[28] V.H. De la Peña and E. Giné. *Decoupling: From Dependence to Independence*. Springer New York, 1999.

[29] J. Dong, W. Wang, and W. Dai. Analysis SimCO: A new algorithm for analysis dictionary learning. In *ICASSP14*, 2014.

[30] D.L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4): 1289–1306, 2006.

[31] D.L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4): 1289–1306, 2006.

[32] 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.

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

[34] 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.

[35] S. Foucart and H. Rauhut. *A Mathematical Introduction to Compressive Sensing*. Applied and Numerical Harmonic Analysis. Birkhäuser, 2013.

[36] J. Fuchs. On sparse representations in arbitrary redundant bases. *IEEE Transactions on Information Theory*, 50:1341–1344, 2004.

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

[38] 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.

[39] 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.

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

[41] Standford ML group. MRNet Dataset. `https://stanfordmlgroup.github.io/competitions/mrnet/`, 2019. Accessed: 2022-06-27.

[42] J. Hajek. Asymptotic theory of rejective sampling with varying probabilities from a finite population. *Annals of Mathematical Statistics*, 35(4):1491–1523, 1964.

[43] O. Hirzallah and F. Kittaneh. Inequalities for sums and direct sums of hilbert space operators. *Linear Algebra and its Applications*, 424(1):71–82, 2007.

[44] R.A. Horn and C.R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge; New York, 2nd edition, 2013.

[45] K. Jogdeo and S. M. Samuels. Monotone convergence of binomial probabilities and a generalization of ramanujan's equation. *Ann. Math. Statist.*, 39:1191–1195, 1968.

[46] F. Krahmer and R. Ward. Stable and robust sampling strategies for compressive imaging. *arXiv:1210.2380*, 2012.

[47] 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.

[48] P. Kuppinger, G. Durisi, and H. Bolcskei. Uncertainty relations and sparse signal recovery for pairs of general signal sets. *IEEE Transactions on Information Theory*, 58:263–277, 2012.

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

[50] 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.

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

[52] S. Minsker. On some extensions of Bernstein's inequality for self-adjoint operators. *Statistics and Probability Letters*, 127:111–119, 2017.

[53] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM J. Comput.*, 24 (2):227–234, 1995.

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

[55] Y. Nesterov. Smooth minimization of nonsmooth functions. *Math. Programming*, pages 127–152, 2005.

[56] M. Pali and K. Schnass. Dictionary learning – from local towards global and adaptive. 2018.

[57] M. Pali, S. Ruetz, and K. Schnass. Average performance of omp and thresholding under dictionary mismatch. *IEEE Signal Processing Letters*, 29:1077–1081, 2022.

[58] M.-C. Pali. *Dictionary Learning & Sparse Modelling*. PhD thesis, University of Innsbruck, 2021.

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

[60] G. Puy, P. Vandergheynst, and Y. Wiaux. On variable density compressive sampling. *IEEE Signal Processing Letters*, 18(10):595–598, 2011.

[61] Q. Qu, Y. Zhai, X. Li, Y. Zhang, and Z. Zhu. Geometric analysis of nonconvex optimization landscapes for overcomplete learning. In *International Conference on Learning Representations*, 2020.

[62] P. Randall. *Sparse recovery via convex optimization*. Ph.d. thesis, n.4349, California Institute of Technology, 2009.

[63] H. Rauhut. Compressive sensing and structured random matrices. In M. Fornasier, editor, *Theoretical Foundations and Numerical Methods for Sparse Recovery*, Radon Series on Computational and Applied Mathematics. De Gruyter Verlag, 2010.

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

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

[66] S. Ruetz. Adapted variable density subsampling for compressed sensing. 2022. doi: 10.48550/arxiv.2206.13796.

[67] S. Ruetz and K. Schnass. Submatrices with non-uniformly selected random supports and insights into sparse approximation. *SIAM Journal on Matrix Analysis and Applications*, 42(3):1268–1289, 2021.

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

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

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

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

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

[73] K. Schnass. Average performance of Orthogonal Matching Pursuit (OMP) for sparse approximation. *IEEE Signal Processing Letters*, 25(12):1865–1869, 2018.

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

[75] K. Schnass and P. Vandergheynst. Dictionary preconditioning for greedy algorithms. *IEEE Transactions on Signal Processing*, 56(5):1994–2002, 2008.

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

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

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

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

[80] J. Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, October 2004.

[81] J. Tropp. Recovery of short, complex linear combinations via l1 minimization. *IEEE Transactions on Information Theory*, 51:1568–1570, 2005.

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

[83] J. Tropp. Norms of random submatrices and sparse approximation. *Comptes Rendus Mathematique*, 346:1271–1274, 2008.

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