Convergence of alternating minimisation algorithms for dictionary learning

Simon Ruetz   SIMON.RUETZ@UIBK.AC.AT
Karin Schnass   KARIN.SCHNASS@UIBK.AC.AT
Universität Innsbruck
Technikerstraße 13
6020 Innsbruck, Austria

Abstract

In this paper we derive sufficient conditions for the convergence of two popular alternating minimisation algorithms for dictionary learning - the Method of Optimal Directions (MOD) and Online Dictionary Learning (ODL), which can also be thought of as approximative K-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, both algorithms will converge with geometric convergence rate to the generating dictionary. This is done even for data models with non-uniform distributions on the supports of the sparse coefficients. These allow the appearance frequency of the dictionary elements to vary heavily and thus model real data more closely.

Keywords: dictionary learning; sparse matrix factorisation; sparse coding; Method of Optimal Directions; MOD; Online Dictionary Learning; ODL; K-Singular Value Decomposition; K-SVD; convergence; non-uniform support distribution; rejective sampling

1. Introduction

The goal of dictionary learning is to find compact data representations by factorising a data matrix $Y \in \mathbb{R}^{d \times N}$ into the product of a dictionary matrix $\Phi \in \mathbb{R}^{d \times K}$ with normalised columns and a sparse coefficient matrix $X \in \mathbb{R}^{K \times N}$

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

The sparsifying dictionary or sparse components of a data class can be used to identify further structure in the data, [13], or simply be exploited for many data processing tasks, such as signal restoration or compressed sensing, [20, 10, 7]. Following the seminal Bayesian approach, [13], there exist several strategies and algorithms to tackle the above problem [13, 3, 12, 35, 17, 18, 19, 37, 33] and a growing number of theoretical results to accompany them [14, 38, 4, 34, 15, 6, 5, 1, 2, 39, 40, 8, 36, 26, 23]. These fall mainly into two categories - optimisation based approaches and graph clustering algorithms. While graph clustering algorithms have stronger theoretical success guarantees, especially in the overcomplete case $K > d$, [2, 4], optimisation based approaches, in particular alternating optimisation algorithms, are extremely successful and popular in practice. The common starting point of three golden classics among these alternating optimisation algorithms,

1

the Method of Optimal Directions (MOD), [12], the K-Singular Value Decomposition (K-SVD), [3] and the algorithm for Online Dictionary Learning (ODL), [19], is the following programme. Given the data matrix $Y = (y_1, \ldots, y_N)$ and a dictionary size K find a dictionary $\Psi = (\psi_1, \ldots, \psi_K)$ and sparse coefficients $X = (x_1, \ldots x_N)$ that optimise

$$\mathrm{argmin}_{\Psi, X} \|Y - \Psi X\|_F^2 \qquad \text{s.t.} \quad X \in \mathcal{S} \quad \text{and} \quad \|\psi_k\|_2 = 1 \quad \text{for all } k. \qquad (2)$$

The set $\mathcal{S}$ enforces sparsity on the coefficients matrix, for instance by allowing only S non-zero coefficients per column, that is $\|x_n\|_0 \leq S$. The norm constraint on the dictionary elements, also called atoms, removes the scaling ambiguity between the dictionary and the coefficients. Note that there remains a sign ambiguity, so for any solution to the problem above, we get 2^K equivalent solutions by flipping the signs of the atoms and adjusting X accordingly. Since the problem is not convex there might exist even more local and/or global minima and to find a local minimum one cannot use gradient descent as the gradient with respect to (Ψ, X) does not have a closed form solution. This is why the three classic algorithms from above employ alternating optimisation and alternate between optimising the coefficients while keeping the dictionary fixed and optimising the dictionary while keeping the coefficients fixed. To update the coefficients, one aims for instance to find

$$\hat{X} = \mathrm{argmin}_X \|Y - \Psi X\|_F^2 = \mathrm{argmin}_X \sum_n \|y_n - \Psi x_n\|_2^2 \qquad \text{s.t.} \quad \|x_n\|_0 \leq S, \qquad (3)$$

which corresponds to N sparse approximation problems. While solving these problems exactly is NP-complete in general, [22, 41], and thus unfeasible for large S, there exist many efficient routines, which have success guarantees under additional conditions and perform well in practice, such as OMP, [25], used for MOD and K-SVD. LARS, [11], used for ODL solves a related problem. In this paper we consider simple thresholding, which is on par with computationally more involved algorithms like OMP in a dictionary learning context, [24], and still relatively easy to analyse theoretically.
Conversely, to update the dictionary for fixed coefficients $\hat{X}$ one aims to find

$$\mathrm{argmin}_{\Psi} \|Y - \Psi \hat{X}\|_F^2 =: f(\Psi) \qquad \text{s.t.} \quad \|\psi_k\|_2 = 1. \qquad (4)$$

The approach leading to the MOD update, [12], is to give up the unit norm constraint on the dictionary elements in (4). The modified problem then has a closed form solution, $\mathrm{argmin}_{\Psi} f(\Psi) = Y\hat{X}^\dagger$, so one can simply update the dictionary as $Y\hat{X}^\dagger D$, where D is a diagonal matrix ensuring that each atom has unit norm.
Another idea is to solve (4) approximately with one step of projected block coordinate descent. So we first calculate the gradient of f with respect to Ψ resulting in

$$\nabla_{\Psi} f(\Psi) = -\sum (y_n - \Psi \hat{x}_n)\hat{x}_n^* = -Y\hat{X}^* + \Psi\hat{X}\hat{X}^*.$$

We then choose adapted step sizes for each dictionary element stored in the diagonal matrix Λ and – with D again a diagonal matrix ensuring normalisation – update the dictionary as

$$\left[\Psi + (Y\hat{X}^* - \Psi\hat{X}\hat{X}^*) \cdot \Lambda\right] \cdot D = \left[\Psi\Lambda^{-1} - Y\hat{X}^* + \Psi\hat{X}\hat{X}^*\right] \cdot \Lambda D. \qquad (5)$$

Algorithm 1.1: MOD and ODL (aK-SVD) - one iteration

Input : Ψ, Y, S, κ (with $\kappa > 1$)

Set $\hat{X} = (\hat{x}_1, \ldots, \hat{x}_N) = 0$

foreach n **do**

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

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

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

$\qquad\qquad \hat{x}_n \leftarrow 0$ `// set pathological estimates to zero`

$\qquad$ **end**

end

$\Psi \leftarrow \begin{cases} Y\hat{X}^\dagger & \text{\texttt{// MOD update}} \\ Y\hat{X}^* - \Psi\hat{X}\hat{X}^* + \Psi \, \text{diag}(\hat{X}\hat{X}^*) & \text{\texttt{// ODL (aK-SVD) update}} \end{cases}$

$\Psi \leftarrow (\psi_1/\|\psi_1\|_2, \ldots, \psi_K/\|\psi_K\|_2)$ `// atom normalisation`

Output: Ψ

The particular choice $\Lambda^{-1} = \text{diag}(\hat{X}\hat{X}^*)$ leads to the dictionary update in ODL, [19].
Finally, for K-SVD the update of the k-th atom and corresponding coefficients can be derived from (4) or rather (2) by fixing $\Psi, \hat{X}$ except for the k-th atom ψ_k and its corresponding locations (but not values) of non-zero entries in the k-th row of $\hat{X}$ and optimising for both using a singular value decomposition. For more details on the update we refer to [3]. However, we want to point out that K-SVD and ODL can be seen as closely related. Indeed if we take an approximate version of K-SVD, known as aK-SVD, [27], where the SVD is replaced with a 1 step power iteration, and additionally skip the coefficient update during the dictionary update, we arrive again at ODL, see [29] for more details. To keep the close link in mind we will refer to the dictionary update as in (5) as ODL (aK-SVD) update.
As already mentioned the disadvantage of the practically efficient, alternating optimisation algorithms for dictionary learning is that they are less supported by theory. In particular, when assuming that the data follows a random sparse model with an overcomplete generating dictionary Φ, there are no global recovery guarantees, in the sense that one of the algorithms above or a variant will recover Φ with high probability.

Our Contribution In this paper we will cement the theoretical foundations of dictionary learning via alternating minimisation by characterising the convergence basin for both MOD and ODL (aK-SVD) in combination with thresholding as sparse approximation algorithm as summarised in Algorithm 1. In particular we will show that under mild conditions on the generating dictionary, any initialisation within atom-wise ℓ_2-norm and scaled operator norm distance of order $1/\log K$ will converge exponentially close to the generating dictionary given enough training signals. Moreover, we provide further conditions on an initialisation that ensure convergence even for distances close to the maximal distance $\sqrt{2}$.
In case of ODL (aK-SVD) this is the first result of this kind. For MOD a convergence radius of order $1/S^2$ was derived in [1], however, our results show that even under milder assumptions the basin of convergence is several orders of magnitude larger. Finally, our most important contribution is that we provide the first theoretical results which are valid

for data generating models where the non-zero entries of the generating coefficients are not essentially uniformly distributed. This means that some dictionary elements can be more likely to appear in the data representation than others. Such a non-homogeneous use of dictionary elements can usually be observed in dictionaries learned on image or audio data. Our results, proving the stability and convergence of alternating minimisation methods for non-uniform sparse support models, might be an explanation for their practical advantages over graph clustering or element-wise dictionary learning methods.

Organisation The remainder of the paper is organised as follows. After introducing the necessary notation in Section 2 we will define our random signal model and give some intuition why MOD and ODL should converge in Section 3. We then provide our main result together with some explanations and connections to other work and the proof of the main theorem in Section 4. However, to make the proof more accessible, the four lemmas it relies on are deferred to Appendix A and further technical prerequisites to Appendix B. We conclude with a discussion of our results and an outline of future work in Section 5.

2. Notation and setting

The notation and setting follows [32, 29] very closely. Let $A \in \mathbb{R}^{d \times K}$ and $B \in \mathbb{R}^{K,m}$. Let A_k and A^k denote the k-th column and k-th row of A respectively and A^* denote the transpose of the matrix A. For $1 \leq p, q, r \leq \infty$ we set $\|A\|_{p,q} := \max_{\|x\|_q=1} \|Ax\|_p$. Recall that $\|AB\|_{p,q} \leq \|A\|_{q,r} \|B\|_{r,p}$ and $\|Ax\|_q \leq \|A\|_{q,p} \|x\|_p$. Often encountered quantities are $\|A\|_{2,1} = \max_{k \in \{1,...,K\}} \|A_k\|_2$ and $\|A\|_{\infty,2} = \max_{k \in \{1,...,d\}} \|A^k\|_2$, which denote the maximal ℓ_2 norm of a column resp. row of A. For ease of notation we sometimes write $\|A\| = \|A\|_{2,2}$ for the largest absolute singular value of A. For a vector $v \in \mathbb{R}^d$, we denote by $\underline{v} := \min_i |v_i|$ the smallest absolute value of v and $\|v\|_\infty$ the maximal entry of v. For a subset $I \subseteq \mathbb{K} := \{1, \ldots, K\}$, called the support, we denote by A_I the submatrix with columns indexed by I and $A_{I,I}$ the submatrix with rows and columns indexed by I. 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 use the convention that subscripts take precedent over transposing, e.g. $A_I^* = (A_I)^*$. We denote by $\mathbf{1}_I \in \mathbb{R}^K$ the vector, whose entries indexed by I are 1 and zero else. Further, for any vector v we denote by D_v resp. $\text{diag}(v)$ 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/vectors of the same dimension.

Throughout this paper we will denote by $\Phi \in \mathbb{R}^{d \times K}$ the generating dictionary, i.e. the ground truth we want to recover, and by $\Psi \in \mathbb{R}^{d \times K}$ the current guess. Wlog we will assume that the columns of Ψ are signed in way that the vector $\alpha \in \mathbb{R}^K$ defined via $\alpha_i := \langle \phi_i, \psi_i \rangle$ has only positive entries. We will denote the ℓ_2-distance between dictionary elements by

$$\varepsilon(\Psi, \Phi) := \|\Psi - \Phi\|_{2,1} = \max_i \|\psi_i - \phi_i\|_2 \quad \Leftrightarrow \quad \varepsilon(\Psi, \Phi)^2 = 2 - 2\underline{\alpha}. \qquad (6)$$

If it is clear from context, we will sometimes write ε instead of $\varepsilon(\Psi, \Phi)$. An important variable which will be used frequently throughout this paper is $Z := \Psi - \Phi$ — the difference matrix between the generating dictionary Φ and a guess Ψ. We define the distance between

4

Φ and Ψ as

$$\delta(\Psi, \Phi) := \max\left\{\|(\Psi - \Phi)D_{\sqrt{\pi}}\|, \|\Psi - \Phi\|_{2,1}\right\} = \max\left\{\|ZD_{\sqrt{\pi}}\|, \|Z\|_{2,1}\right\}. \qquad (7)$$

This might not seem intuitive at first glance, but to show convergence we have to control the weighted operator norm of the difference matrix as well as the ℓ_2-distance. Again, if it is clear from context we will simply write δ.

3. Probabilistic model

As was noted in the introduction we want the locations of the non-zero coefficients to follow a non-uniform distribution, allowing some dictionary elements to be picked more frequently than others.

Definition 1 (Poisson and rejective sampling) *Let δ_k denote a sequence of K independent Bernoulli 0-1 random variables with expectations $0 \leq p_k \leq 1$ such that $\sum_{k=1}^{K} p_k = S$ and denote by $\mathbb{P}_B$ the probability measure of the corresponding Poisson sampling model. We say the support I follows the Poisson sampling model, if $I := \{k \mid \delta_k = 1\}$ and each support $I \subseteq \mathbb{K}$ is chosen with probability*

$$\mathbb{P}_B(I) = \prod_{i \in I} p_i \prod_{j \notin I}(1 - p_j). \qquad (8)$$

We say our support I follows the rejective sampling model, if each support $I \subseteq \mathbb{K}$ is chosen with probability

$$\mathbb{P}_S(I) := \mathbb{P}_B(I \mid |I| = S). \qquad (9)$$

If it is clear from the context, we write $\mathbb{P}(I)$ instead of $\mathbb{P}_S(I)$.

The Poisson sampling model would be quite convenient since the probability of one atom appearing in the support is independent of the others. Unfortunately, we need a model with exactly S-sparse supports, rather than supports that are S-sparse on average. The rejective model satisfies this second condition and still yields almost independent atoms, in the sense that we can often reduce estimates for rejective sampling to estimates for Poisson sampling, [32]. Note also that unless we are in the uniform case, where $p_i = S/K$ for all i, the inclusion probabilities $\pi_i := \mathbb{P}(i \in I)$ are different from the parameters p_i. They can be related using [32][Lemma 1], which we restate for convenience in Appendix B.1. We frequently use the square diagonal matrix $D_{\sqrt{\pi}} := \text{diag}((\sqrt{\pi_k})_k)$. Based on the rejective sampling model, we define the following model for our signals.

Definition 2 (Signal model) *Given a generating dictionary $\Phi \in \mathbb{R}^{d \times K}$ consisting of K normalized atoms, we model our signals as*

$$y = \Phi_I x_I = \sum_{i \in I} \phi_i x_i, \quad x_i = c_i \sigma_i, \qquad (10)$$

where the support $I = \{i_1, \ldots i_S\} \subseteq \mathbb{K}$ is chosen according to the rejective sampling model (9) with parameters $p_1, \ldots, p_K$ such that $\sum_{i=1}^{K} p_i = S$ and $0 < p_k \leq 1/6$, the coefficient sequence

$c = (c_i)_i \in \mathbb{R}^K$ consists of i.i.d. bounded random variables c_i with $0 \le c_{\min} \le c_i \le c_{\max} \le 1$ 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 $p_i \le 1/6$ ensures that the p_i and the corresponding inclusion probabilities of the rejective sampling model π_i are not too different (see Theorem 15). We further introduce the vector $\beta \in \mathbb{R}^K$ via $\beta_i := \mathbb{E}[c_i^2]$ and denote by D_β the corresponding diagonal matrix. To see how this signal model allows us to prove convergence of the algorithms, note that before normalisation, by using $Y = \Phi X$, we can write the two dictionary update steps concisely as

MOD: $\quad Y\hat{X}^*(\hat{X}\hat{X}^*)^{-1} = \Phi X\hat{X}^*(\hat{X}\hat{X}^*)^{-1}$,

ODL: $\quad \dfrac{1}{N}\left[Y\hat{X}^* - \Psi\hat{X}\hat{X}^* + \Psi\,\mathrm{diag}(\hat{X}\hat{X}^*)\right] = \dfrac{1}{N}\left[\Phi X\hat{X}^* - \Psi\hat{X}\hat{X}^* + \Psi\,\mathrm{diag}(\hat{X}\hat{X}^*)\right],$

where we scaled the update step of the ODL algorithm by a factor of $1/N$. The key is to show that both these update steps concentrate around Φ. To that end we define the two averages of random matrices

$$A := \frac{1}{N}X\hat{X}^* = \frac{1}{N}\sum_{n=1}^{N} x_n\hat{x}_n^* \quad \text{and} \quad B := \frac{1}{N}\hat{X}\hat{X}^* = \frac{1}{N}\sum_{n=1}^{N}\hat{x}_n\hat{x}_n^*. \qquad (11)$$

With these we can write the update step of MOD as ΦAB^{-1} and that of ODL as $\Phi A - \Psi[B - \mathrm{diag}(B)]$. 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 I, 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 approximate $\Psi_I^\dagger \approx \Psi_I^*$, leading to

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

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,

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

So the empirical estimator $A = \frac{1}{N}\sum_{n=1}^{N} x_n\hat{x}_n^* \approx \mathbb{E}[x\hat{x}]$ will be well approximated by

$$\mathbb{E}[x\hat{x}^*] \approx \mathbb{E}\,\mathrm{diag}(\mathbf{1}_I)D_\beta\Phi^*\Psi\,\mathrm{diag}(\mathbf{1}_I) = (D_\beta\Phi^*\Psi) \odot \mathbb{E}[\mathbf{1}_I\mathbf{1}_I^*].$$

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 are 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 $D_\alpha = \mathrm{diag}(\Phi^*\Psi)$ a similar analysis for B yields

$$A \approx \mathbb{E}[x\hat{x}^*] \approx (D_\beta\Phi^*\Psi) \odot D_\pi = D_{\pi\cdot\alpha\cdot\beta} \quad \text{and} \quad B \approx \mathbb{E}[\hat{x}\hat{x}^*] \approx D_{\pi\cdot\alpha^2\cdot\beta}.$$

So before normalisation the updates via MOD and ODL should be approximately

$$\Phi A B^{-1} \approx \Phi D_\alpha^{-1} \quad \text{and} \quad \Phi A - \Psi[B - \text{diag}(B)] \approx \Phi D_{\pi \cdot \alpha \cdot \beta}. \tag{12}$$

This means that the output of both dictionary update steps after normalisation should be very close to the ground truth, and the proof boils down to quantifying the error in the approximation steps outlined above.

4. Main result

Concretely, we will prove the following theorem.

Theorem 3 *Assume our signals follow the signal model in 2. Define*

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

with $\kappa^2 \geq 2$ and let C, n be two universal constants, where C is no larger than 42 and n no larger than 130. Denote by $\delta_\star$ the desired recovery accuracy and assume $\delta_\star \log(nK\rho/\delta_\star) \leq \gamma/C$. We abbreviate $\nu = 1/\sqrt{\log(nK\rho/\delta_\star)} < 1$. If the atom-wise distance $\varepsilon = \varepsilon(\Psi, \Phi)$ of the current guess Ψ to the generating dictionary Φ satisfies

$$\max\left\{ \nu \|\Phi D_{\sqrt{\pi}}\|, \mu(\Phi) \right\} \leq \left(1 - \frac{\varepsilon^2}{2} \right) \cdot \frac{\gamma}{4C \log(nK\rho/\delta_\star)} \tag{13}$$

and the current guess Ψ additionally satisfies either

$$\max\left\{ \nu \|\Psi D_{\sqrt{\pi}}\|, \mu(\Psi), \mu(\Psi, \Phi) \right\} \leq \left(1 - \frac{\varepsilon^2}{2} \right) \cdot \frac{\gamma}{4C \log(nK\rho/\delta_\star)} \tag{14}$$

$$\text{or} \qquad \delta(\Psi, \Phi) \leq \frac{\gamma}{C \log(nK\rho/\delta_\star)} =: \delta_\circ, \tag{15}$$

then the updated and normalised dictionary $\hat{\Psi}$, which is output by ODL or MOD, satisfies

$$\delta(\hat{\Psi}, \Phi) \leq \frac{1}{2} \cdot \left(\delta_\star/2 + \min\{\delta_\circ, \delta(\Psi, \Phi)\} \right) =: \frac{1}{2} \cdot \Delta, \tag{16}$$

except with probability

$$60K \exp\left(-\frac{N(\Delta/16)^2}{2\rho^2 + \rho\Delta/16} \right). \tag{17}$$

To make the theorem more accessible, we now provide a detailed discussion of the implications and assumptions.

Convergence First note that a repeated application of the theorem proves convergence of MOD and ODL (aK-SVD) up to accuracy $\delta_\star$, assuming a generating dictionary and an initialisation that satisfies the conditions above and given a new batch of training signal in each iteration of size which scales at worst as $N \approx \log K\rho^2/\delta_\star^2$. To see this, note that the theorem ensures that the distance between the generating dictionary and the output

dictionary decreases fast enough for condition (15) to be satisfied after one step. Once an input dictionary satisfies $\delta_\star \leq \delta(\Psi, \Phi) \leq \delta_\circ$ each iteration will w.h.p. shrink the distance by a factor $\eta \leq 3/4$, since we have

$$\delta(\hat{\Psi}, \Phi) \leq \tfrac{1}{4} \cdot \delta_\star + \tfrac{1}{2} \cdot \delta(\Psi, \Phi) \leq \tfrac{3}{4} \cdot \delta(\Psi, \Phi).$$

As the conditions on the generating dictionary get easier to satisfy in each step, we can iterate this argument to get convergence.

Finally, if an input dictionary already satisfies $\delta(\Psi, \Phi) \leq \delta_\star$, that is, we reached the desired precision, the contraction property (16) guarantees that the output dictionary still satisfies $\delta(\hat{\Psi}, \Phi) \leq \tfrac{3}{4}\delta_\star \leq \delta_\star$. This stability is due to the assumption that the target recovery accuracy satisfies $\delta_\star \log (nK\rho/\delta_\star) \leq \gamma/C$ meaning $\delta_\star \leq \delta_\circ$. Going through the proof shows that without this assumption, for instance because there are not enough training signals for a higher accuracy available, the theorem above is still true and the target distance will be achieved after one step. In this case however, since the output dictionary might no longer satisfy (14), yet not be close enough to satisfy (15), it is not guaranteed that a further iteration will preserve this distance.

Generating dictionary The first condition in (13) contains as implicit conditions on the generating dictionary that

$$\|\Phi D_{\sqrt{\pi}}\| \lesssim \frac{1}{\sqrt{\log K}} \quad \text{and} \quad \mu(\Phi) \lesssim \frac{1}{\log K}, \tag{18}$$

which together ensure that Φ is a sensible dictionary. It guarantees that most random sparse supports are well-conditioned, that is $\|\Phi_I^*\Phi_I - \mathbb{I}\| \leq \vartheta < 1$. This further means that most signals generated by the model have a stable representation, $y = \Phi_I x_I$ with $\|y\|_2 \approx \|x_I\|_2$, and could be identified by a sparse approximation algorithm that has access to Φ. To see that the condition is quite mild, note that for uniformly distributed supports, $p_i = \pi_i = S/K$, and Φ a unit norm tight frame, $\|\Phi\| = K/d$, we can rewrite it as $S \leq d/\log K$, which is for instance a common requirement in compressed sensing.

Initialisation (input dictionary): The condition in (13) further limits the maximal atom-wise distance of an initialisation. Assuming again uniformly distributed supports and a tight generating dictionary it corresponds to

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

Considering that the maximal distance between two unit norm vectors is $\sqrt{2}$, this means that convergence is possible even from far away initialisations. However, to actually converge we have additional requirements. The initialisation needs to be well behaved, which is ensured by (14) via $\|\Psi D_{\sqrt{\pi}}\| \lesssim \frac{1}{\sqrt{\log K}}$ and $\mu(\Psi) \lesssim \frac{1}{\log K}$, similar to (18) for the generating dictionary. The more stringent and interesting requirement in (14) is the condition on the cross-coherence $\mu(\Psi, \Phi)$ or rather the cross-Gram matrix $\Psi^*\Phi$. As the minimal entry on its diagonal is $\underline{\alpha} = (1 - \varepsilon^2/2)$ it translates to

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

or $\Psi^*\Phi$ being diagonally dominant. Intuitively, (13) and (14) mean that the admissible distance can be very close to $\sqrt{2}$, as long as the initialisation is a well-behaved dictionary and no two estimated atoms point to the same generating atom, meaning it is clear to the sparse approximation algorithm which estimated atom belongs to which generating atom. While it might be possible to relax this separation condition, it is unlikely that it can be removed in the overcomplete case. Indeed, following the guidelines in [23], one can construct well-behaved incoherent initialisations for which both MOD and ODL converge to a local minimum that is not equivalent to the generating dictionary.

Finally, if the initialisation is within distance $1/\log K$ to the generating dictionary, meaning condition (15) is satisfied, it is automatically well-behaved and has a diagonally dominant cross-Gram matrix, because it inherits these properties from the generating dictionary, which by (13) is well behaved and incoherent.

Number of signals: From the probability bound in (17) 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

$$N \approx \frac{\rho^2}{\delta_\star^2} \cdot \log K \approx \frac{1}{\underline{\pi}^3 \delta_\star^2} \cdot \log K.$$

This ensures that even the most rarely appearing atoms are seen often enough to learn them properly. The relation above reflects the general dependencies, meaning we need more training signals for higher accuracy and more imbalanced atom distributions, but is a little too pessimistic. We expect that the scaling can be reduced to $N \approx \log K \delta_\star^{-1} \underline{\pi}^{-3/2}$ and even $N \approx K \log K / \delta_\star$ in the uniform case, but leave the endeavour to those still motivated after reading the current proof[1].

Attainable accuracy: We have already seen that the target accuracy should not be chosen too large in order to have stable convergence and that the number of training signals required in each iteration should grow with the desired accuracy. Another interesting observation is that even given an arbitrarily close initialisation and arbitrarily many training signals, the best attainable accuracy is limited by the coherence and conditioning of the generating dictionary via (13) — even in the noiseless case considered here. The main reason is that (13) is a very light condition which does not exclude the existence and rare selection of sparse supports, leading to ill-conditioned or even rank-deficient matrices Ψ_I. Such supports cannot be recovered by thresholding or any other sparse approximation algorithm. This failure probability stops the algorithm from attaining arbitrarily small precision. If we exchange (13) by the more restrictive assumption $2S\mu(\Phi) < 1$ then all supports of size S are well enough conditioned to be identified by a sparse algorithm like OMP using Φ (or some very small perturbation of it). This means that Φ is a fixed point of the algorithm and that for a close enough initialisation we have convergence to Φ in expectation.

Comparison to existing results for MOD: Finally, we can compare our result with that in [1] for MOD with an ℓ_1-minimisation based sparse approximation routine. Assuming a signal model with uniformly selected random supports and a non-zero coefficient distribution, which is bounded from above, it is shown that for an approximately tight dictionary with coherence $\mu(\Phi) \lesssim 1/\sqrt{d}$ and a sparsity level $S \lesssim d^{1/6}$ MOD will converge from

---

1. Hints how to proceed can be found after the proof Lemma 10

any initialisation satisfying $\varepsilon(\Psi, \Phi) \lesssim 1/S^2$, given enough training signals. In this special case our assumptions on the generating dictionary — $\mu(\Phi) \lesssim 1/\log K$ and $S \lesssim d/\log(K)$ — are more relaxed. As already seen, this comes at the price of a theoretical limit on the achievable accuracy, which in practice, however, is determined by the number of available signals. Further, we also have lighter conditions on the distance between the initialisation and the generating dictionary. In particular, for a tight dictionary our conditions can be written as

$$\varepsilon(\Psi, \Phi) \lesssim \frac{1}{\log K} \quad \text{and} \quad \|\Psi - \Phi\| \lesssim \sqrt{\frac{K}{S \log^2 K}},$$

whereas the assumption in [1] that $\varepsilon(\Psi, \Phi) \lesssim 1/S^2$ implies $\|\Psi - \Phi\| \lesssim \sqrt{K}/S^2$ by using $\|A\| \leq \|A\|_F \leq \sqrt{K}\|A\|_{2,1}$. Hence the assumptions in [1] are more restrictive.

These restrictions seem due to using ℓ_1-minimisation for sparse approximation, which on the other hand has the advantage that it does not require the distribution of the non-zero coefficients to be bounded away from zero. However, such an assumption is used in the graph clustering algorithm suggested in [1] to get an initialisation, that satisfies the convergence conditions. This indicates that such a condition is necessary to get a larger convergence area.

Proof [Proof of Theorem 3] We first collect the results of Lemmas 7-10. Writing

$$T := (D_{\sqrt{\pi} \cdot \alpha})^{-1} B (D_{\sqrt{\pi} \cdot \alpha \cdot \beta})^{-1} \quad \text{and} \quad \mathbb{I}_{\ell^c} := \mathbb{I} - e_\ell e_\ell^* \tag{19}$$

for convenience, we get that except with failure probability as in (17) we have

$$\|\Phi A (D_{\sqrt{\pi} \cdot \alpha \cdot \beta})^{-1} - \Phi D_{\sqrt{\pi}}\|_{2,2} \leq \underline{\alpha} \Delta/8 \qquad \text{and} \qquad \|T - \mathbb{I}\|_{2,2} \leq \Delta/4,$$

as well as for all $\ell \in \{1, \cdots, K\}$

$$\|\Phi A (D_{\pi \cdot \alpha \cdot \beta})^{-1} e_\ell - \phi_\ell\|_2 \leq \frac{\Delta}{8} \quad \text{and} \quad \max\left\{\frac{\gamma \underline{\alpha} \nu}{4C}, \|\Psi D_{\sqrt{\pi}}\|\right\} \cdot \|\mathbb{I}_{\ell^c} T e_\ell \pi_\ell^{-1/2}\|_2 \leq \frac{\Delta}{8}.$$

Using these 4 inequalities we show that for both algorithms a properly scaled version of the updated dictionary, which we denote by $\bar{\Psi}$, contracts towards the generating dictionary. Concretely, we show that for some constants s_ℓ close to 1

$$\|(\bar{\Psi} - \Phi) D_{\sqrt{\pi}}\| \leq \Delta/4 \quad \text{and} \quad \max_\ell \|\bar{\psi}_\ell - s_\ell \phi_\ell\| \leq \Delta/3. \tag{20}$$

Together these bounds guarantee that the normalised version of the updated dictionary $\hat{\Psi}$ satisfies $\delta(\hat{\Psi}, \Phi) \leq \Delta/2$, meaning we have contraction towards the generating dictionary in the weighted operator norm and the maximum column norm simultaneously. We start with the proof of ODL which is a bit simpler.

ODL: Motivated by 12, we define a scaled version of the updated dictionary $\bar{\Psi} := [\Phi A - \Psi B + \Psi \operatorname{diag}(B)] (D_{\pi \cdot \alpha \cdot \beta})^{-1}$, which ensures that on average $\bar{\Psi}$ concentrates around

Φ. The scaling does not change the underlying algorithm, since we have a normalisation step at the end of each iteration, which we will analyse afterwards. We decompose $\bar{\Psi}$ as

$$
\begin{aligned}
\bar{\Psi} &= \Phi A (D_{\pi \cdot \alpha \cdot \beta})^{-1} - \Psi [B - \operatorname{diag}(B)](D_{\pi \cdot \alpha \cdot \beta})^{-1} \\
&= \Phi A (D_{\pi \cdot \alpha \cdot \beta})^{-1} - \Psi D_{\sqrt{\pi} \cdot \alpha} \left[(D_{\sqrt{\pi} \cdot \alpha})^{-1} B (D_{\pi \cdot \alpha \cdot \beta})^{-1} - D_{\sqrt{\pi}}^{-1} \right] \\
&\qquad\qquad\qquad + \Psi D_{\sqrt{\pi} \cdot \alpha} \left[(D_{\sqrt{\pi} \cdot \alpha})^{-1} \operatorname{diag}(B) (D_{\pi \cdot \alpha \cdot \beta})^{-1} - D_{\sqrt{\pi}}^{-1} \right] \\
&= \Phi A (D_{\pi \cdot \alpha \cdot \beta})^{-1} - \Psi D_{\sqrt{\pi} \cdot \alpha} \left(T - \mathbb{I} \right) D_{\sqrt{\pi}}^{-1} + \Psi D_{\sqrt{\pi} \cdot \alpha} \left(\operatorname{diag}(T) - \mathbb{I} \right) D_{\sqrt{\pi}}^{-1}. \tag{21}
\end{aligned}
$$

We first show contraction in the weighted operator norm. Note that in both regimes we have $\|\Psi D_{\sqrt{\pi}}\| \le 2/C$, either by direct assumption or based on the bound

$$
\|\Psi D_{\sqrt{\pi}}\| \le \|\Phi D_{\sqrt{\pi}}\| + \|(\Psi - \Phi) D_{\sqrt{\pi}}\| \le \|\Phi D_{\sqrt{\pi}}\| + \delta \le \|\Phi D_{\sqrt{\pi}}\| + \delta_\circ.
$$

With the expression for the updated dictionary $\bar{\Psi}$ in (21) and using the fact that $\|D_\alpha\| \le 1$ we can bound the operator norm of the difference $(\bar{\Psi} - \Phi) D_{\sqrt{\pi}}$ as

$$
\|(\bar{\Psi} - \Phi) D_{\sqrt{\pi}}\| \le \underbrace{\|\Phi A (D_{\sqrt{\pi} \cdot \alpha \cdot \beta})^{-1} - \Phi D_{\sqrt{\pi}}\|}_{\le \, \alpha \Delta / 8 \; (7)} + \underbrace{2 \|\Psi D_{\sqrt{\pi} \cdot \alpha}\|}_{\le \, 4/C} \cdot \underbrace{\|T - \mathbb{I}\|}_{\cdot \Delta/4 \; (8)} \le \frac{\Delta}{4}.
$$

Next we show that for each atom of the scaled dictionary the ℓ_2-distance also decreases with each iteration. We access the ℓ-th dictionary atom $\bar{\psi}_\ell$ simply by multiplying $\bar{\Psi}$ with the standard basis vector e_ℓ. This yields

$$
\begin{aligned}
\bar{\psi}_\ell = \bar{\Psi} e_\ell &= \Phi A (D_{\pi \cdot \alpha \cdot \beta})^{-1} e_\ell - \Psi D_{\sqrt{\pi} \cdot \alpha} \left[T - \operatorname{diag}(T) \right] D_{\sqrt{\pi}}^{-1} e_\ell \\
&= \Phi A (D_{\pi \cdot \alpha \cdot \beta})^{-1} e_\ell + \Psi D_{\sqrt{\pi} \cdot \alpha} \, \mathbb{I}_{\ell^c} T \, e_\ell \pi_\ell^{-1/2}. \tag{22}
\end{aligned}
$$

Using this decomposition together with our second set of inequalities we get

$$
\|\bar{\psi}_\ell - \phi_\ell\| \le \underbrace{\|\Phi A (D_{\pi \cdot \alpha \cdot \beta})^{-1} e_\ell - \phi_\ell\|}_{\le \, \Delta/8 \; (9)} + \underbrace{\|\Psi D_{\sqrt{\pi} \cdot \alpha}\| \cdot \|\mathbb{I}_{\ell^c} T e_\ell \pi_\ell^{-1/2}\|}_{\le \, \Delta/8 \; (10)} \le \frac{\Delta}{4},
$$

which shows the second part in (20) for $s_\ell = 1$. To finish the proof we still need to show that (20) guarantees contraction in the weighted operator norm and in the maximum column norm after normalisation. However, 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 the second of our 4 inequalities, we can write the dictionary update step before normalisation as $\Phi X \hat{X}^* (\hat{X}\hat{X}^*)^{-1} = \Phi A B^{-1}$. This dictionary update step — though conceptually very easy — is harder to analyse theoretically due to the inverse of the matrix B. Again we will do the analysis for a scaled version of the updated dictionary $\bar{\Psi} := \Phi A B^{-1} D_\alpha \approx \Phi$. As for ODL we start by showing that the weighted operator norm of the difference $\bar{\Psi} - \Phi$ contracts. We split $\bar{\Psi} D_{\sqrt{\pi}}$ as follows

$$
\begin{aligned}
\bar{\Psi} D_{\sqrt{\pi}} &= \Phi A B^{-1} D_{\sqrt{\pi} \cdot \alpha} \\
&= \Phi A (D_{\sqrt{\pi} \cdot \alpha \cdot \beta})^{-1} + \Phi A (D_{\sqrt{\pi} \cdot \alpha \cdot \beta})^{-1} \left(\left[(D_{\sqrt{\pi} \cdot \alpha})^{-1} B (D_{\sqrt{\pi} \cdot \alpha \cdot \beta})^{-1} \right]^{-1} - \mathbb{I} \right) \\
&= \Phi A (D_{\sqrt{\pi} \cdot \alpha \cdot \beta})^{-1} + \Phi A (D_{\sqrt{\pi} \cdot \alpha \cdot \beta})^{-1} \left(T^{-1} - \mathbb{I} \right). \tag{23}
\end{aligned}
$$

Since by Lemma 8 the matrix T is close to the identity, $\|T - \mathbb{I}\| \le \Delta/4$, the Neumann series for its inverse converges and we have $T^{-1} = [\mathbb{I} - (\mathbb{I} - T)]^{-1} = \sum_{k \ge 0} (\mathbb{I} - T)^k$. Using the geometric series formula we get for the operator norms of $T^{-1}, T^{-1} - \mathbb{I}$

$$\|T^{-1}\| = \|\sum_{k \ge 0} (\mathbb{I} - T)^k\| \le \sum_{k \ge 0} \|\mathbb{I} - T\|^k \le \frac{1}{1 - \Delta/4}, \tag{24}$$

$$\|T^{-1} - \mathbb{I}\| = \|\sum_{k \ge 1} (\mathbb{I} - T)^k\| \le \sum_{k \ge 1} \|T - \mathbb{I}\|^k \le \frac{\Delta}{4 - \Delta} \le \frac{\Delta}{3}, \tag{25}$$

where for the last inequality we have used that $\Delta \le \frac{3}{2} \cdot \delta_\circ \le 1/4 \le 1$ since $\delta_\circ = \gamma\nu^2/C$ and $\nu \le 1/3$. Note also that by Lemma 7 and (13), we get

$$\|\Phi A(D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1}\| \le \|\Phi A(D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1} - \Phi D_{\sqrt{\pi}} + \Phi D_{\sqrt{\pi}}\| \le \frac{\alpha\Delta}{8} + \|\Phi D_{\sqrt{\pi}}\| \le \frac{5}{4} \cdot \frac{\alpha\gamma\nu}{4C}. \tag{26}$$

Combining these observations and using the triangle inequality repeatedly yields

$$\|(\bar{\Psi} - \Phi)D_{\sqrt{\pi}}\| \le \underbrace{\|\Phi A(D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1} - \Phi D_{\sqrt{\pi}}\|}_{\le \alpha\Delta/8 \ (7)} + \underbrace{\|\Phi A(D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1}\|}_{\le \alpha\gamma\nu/C \ (26)} \cdot \underbrace{\|T^{-1} - \mathbb{I}\|}_{\le \Delta/3 \ (25)} \le \frac{\Delta}{4}.$$

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

Now to the contraction of the atomwise ℓ_2-norm. First we again split our updated dictionary atom $\bar{\psi}_\ell$ into two parts using that $\mathbb{I} = e_\ell e_\ell^* + \mathbb{I}_{\ell^c}$

$$\bar{\psi}_\ell = \bar{\Psi} e_\ell = \Phi A(D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1} T^{-1} e_\ell \pi_\ell^{-\frac{1}{2}}$$

$$= \Phi A(D_{\pi\cdot\alpha\cdot\beta})^{-1} e_\ell \cdot e_\ell^* T^{-1} e_\ell + \Phi A(D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1} \mathbb{I}_{\ell^c} T^{-1} e_\ell \pi_\ell^{-\frac{1}{2}}. \tag{27}$$

Using the shorthand $s_\ell = e_\ell^* T^{-1} e_\ell$ and substracting $s_\ell \phi_\ell$ we get

$$\|\bar{\psi}_\ell - s_\ell \phi_\ell\| \le |s_\ell| \cdot \|\Phi A(D_{\pi\cdot\alpha\cdot\beta})^{-1} e_\ell - \phi_\ell\| + \|\Phi A(D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1}\| \cdot \|\mathbb{I}_{\ell^c} T^{-1} e_\ell \pi_\ell^{-\frac{1}{2}}\|.$$

The first term is well-behaved and makes no problems. Indeed, since $|s_\ell| \le \|T^{-1}\|$, using (24) and Lemma 9 yields

$$|s_\ell| \cdot \|\Phi A(D_{\pi\cdot\alpha\cdot\beta})^{-1} e_\ell - \phi_\ell\| \le (1 - \Delta/4)^{-1} \cdot \Delta/8. \tag{28}$$

To bound $\|\mathbb{I}_{\ell^c} e_\ell T^{-1} e_\ell\|$ we use $T^{-1} = \mathbb{I} + T^{-1}(\mathbb{I} - T)$ and $\mathbb{I}_{\ell^c} e_\ell = 0$ to get

$$\|\mathbb{I}_{\ell^c} T^{-1} e_\ell\| = \|\mathbb{I}_{\ell^c} T^{-1}(\mathbb{I} - T)e_\ell\| = \|\mathbb{I}_{\ell^c} T^{-1}(e_\ell e_\ell^* + \mathbb{I}_{\ell^c})(\mathbb{I} - T)e_\ell\|$$

$$\le \|\mathbb{I}_{\ell^c} T^{-1} e_\ell\| \cdot \|\mathbb{I} - T\| + \|T^{-1}\| \cdot \|\mathbb{I}_{\ell^c} T e_\ell\|.$$

Rearranging the above and using (24) as well as Lemma 8 yields

$$\|\mathbb{I}_{\ell^c} T^{-1} e_\ell\| \le \frac{\|T^{-1}\|}{1 - \|\mathbb{I} - T\|} \cdot \|\mathbb{I}_{\ell^c} T e_\ell\| \le \frac{1}{(1 - \Delta/4)^2} \cdot \|\mathbb{I}_{\ell^c} T e_\ell\|.$$

Combining these observations with the bound from (26) and Lemma 10 yields

$$\|\bar{\psi}_\ell - s_\ell \phi_\ell\| \le \frac{1}{1 - \Delta/4} \cdot \frac{\Delta}{8} + \frac{1}{(1 - \Delta/4)^2} \cdot \frac{5}{4} \cdot \underbrace{\frac{\alpha\gamma\nu}{4C} \cdot \|\mathbb{I}_{\ell^c} T e_\ell \pi_\ell^{-\frac{1}{2}}\|}_{\le \Delta/8 \quad (10)} \le \frac{\Delta}{3}. \tag{29}$$

12

Normalisation Combining the above results shows that with high probability, the dictionary update step of both algorithms before normalisation satisfies

$$\|(\bar{\Psi} - \Phi)D_{\sqrt{\pi}}\|_{2,2} \leq \Delta/4 \quad \text{and} \quad \max_\ell \|\bar{\psi}_\ell - s_\ell\phi_\ell\| \leq \Delta/3, \tag{30}$$

where $s_\ell = 1$ in case of ODL and $|s_\ell - 1| \leq \|T^{-1} - 1\| \leq \Delta/3$ in case of MOD. So what is left to show is that the normalisation step at the end of each iteration does not interfere with convergence. Let $F := \text{diag}(\|\bar{\psi}_i\|_2)^{-1}$ be the square diagonal normalization matrix and denote by $\hat{\Psi} := \bar{\Psi}F$ the normalized dictionary of the current update step. Since $\|\phi_\ell\| = 1$ we have

$$\big|\|\bar{\psi}_\ell\| - 1\big| \leq \|\bar{\psi}_\ell - \phi_\ell\| \leq \|\bar{\psi}_\ell - s_\ell\phi_\ell\| + \|(s_\ell - 1)\phi_\ell\| \leq \Delta/3 + \Delta/3 \leq \Delta,$$

which further means that $\|F\|_{2,2} \leq (1 - \Delta)^{-1}$ and $\|\mathbb{I} - F\| \leq \Delta \cdot (1 - \Delta)^{-1}$. Hence, using that by assumption $\|\Phi D_{\sqrt{\pi}}\| \leq 1/(4C)$ and $\Delta \leq 2/C$ the weighted operator norm of the difference of the generating dictionary Φ and the normalised update $\hat{\Psi}$ can be bounded as

$$\|(\hat{\Psi} - \Phi)D_{\sqrt{\pi}}\| \leq \|(\bar{\Psi} - \Phi)D_{\sqrt{\pi}}\|\|F\| + \|\Phi D_{\sqrt{\pi}}\|\|(\mathbb{I} - F)\|$$
$$\leq \frac{\Delta}{4} \cdot \frac{1}{1 - \Delta} + \frac{1}{4C} \cdot \frac{\Delta}{1 - \Delta} \leq \frac{\Delta}{2}. \tag{31}$$

To bound the ℓ_2-norm we simply use Lemma B.10 from [35], which says that if $\|\phi\| = 1$ and $\|\psi - s\phi\| \leq t$, then $\hat{\psi} = \psi/\|\psi\|$ satisfies $\|\hat{\psi} - \phi\|^2 \leq 2 - 2\sqrt{1 - t^2/s^2}$. Combining this with the bound $\sqrt{1 - t} \geq 1 - \frac{t}{2-t}$ for $t \in (0, 1)$ and using $t = \Delta/3$ and $|1 - s_\ell| \leq \Delta/3$, meaning $s_\ell \geq 1 - \Delta/3$, yields

$$\|\hat{\psi}_\ell - \phi_\ell\| \leq t \cdot \left(s_\ell^2 - \frac{t^2}{2}\right)^{-1/2} \leq \frac{\Delta}{3} \cdot \left(\left(1 - \frac{\Delta}{3}\right)^2 - \frac{\Delta^2}{18}\right)^{-1/2} \leq \frac{\Delta}{2}.$$

This shows that not only $\|(\hat{\Psi} - \Phi)D_{\sqrt{\pi}}\|$ but also $\|\hat{\Psi} - \Phi\|_{2,1}$ and thus $\delta(\hat{\Psi}, \Phi)$ is bounded by $\Delta/2$ and thus concludes the proof of Theorem 3. ∎

5. Discussion

In this paper we have shown that two widely used alternating minimisation algorithms, MOD and ODL (aK-SVD) both combined with thresholding, converge to a well behaved data-generating dictionary Φ from any initialisation Ψ that either lies within distance $O(1/\log K)$ or is itself well-behaved and has a diagonally dominant cross-coherence matrix $\Psi^*\Phi$. For ODL this constitutes the first convergence result of this kind, while for MOD it extends the convergent areas, derived in [1], by orders of magnitude and under weaker assumptions on the generating dictionary.

We want to emphasize that — to the best of our knowledge — our convergence theorem is the first result in theoretical dictionary learning, which is valid for signal generating models that do not assume a (quasi)-uniform distribution of sparse supports. Instead it can handle more realistic distributions, where each atom is used with a different probability. This stability might also be an explanation for the popularity of alternating minimisation algorithms

in practical applications such as image processing. Natural images, for instance, contain more low than high-frequencies and the success of e.g. wavelets in classical signal processing can be attributed to the fact that they can be divided into frequency bands with different appearance probabilities. Since also more recent schemes such as compressed sensing can by improved by exploiting information about the inclusion/appearance probabilities of each wavelet, [32, 28], we expect that our study of dictionary learning for non-uniform support distributions will encourage the development of exciting new signal processing methods that leverage this additional information about the frequency of appearance (and thus in some sense the importance) of different atoms.

Note that the techniques we used to prove convergence for MOD and ODL here can also be used to turn the contraction results for ITKrM, [23], into convergence results, [30]. As part of our future work we plan to further increase the realism of our data models by including different coefficient distributions for different atoms and noise. We also plan to analyse partial convergence of the dictionary, meaning convergence for most but not all atoms. The main motivation for this is that a randomly initialised dictionary is still rather unlikely to satisfy even the relaxed diagonal dominance requirement. However, we expect that after resorting, a large, left upper part of $\Psi^*\Phi$ will be diagonally dominant and that all associated atoms still converge quite closely. A result of this kind would put the replacement and adaptive dictionary learning strategies, developed in [23], on a firmer theoretical basis. Lastly, we are interested in increasing the realism of our signal model even further by including also second order inclusion probabilities into our generating model. This is again inspired by the behaviour of wavelets in natural images, where it is know that spatially close wavelets are likely to appear together.

Acknowledgments

Appendix A. Proofs of the 4 inequalities

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 set $\vartheta := 1/4$ and write

$$\mathcal{F}_\Phi := \{I \ : \ \|\Phi_I^*\Phi_I - \mathbb{I}\| \leq \vartheta\} \quad \text{and} \quad \mathcal{F}_\Psi := \{I \ : \ \|\Psi_I^*\Psi_I - \mathbb{I}\| \leq \vartheta\} \tag{32}$$

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

$$\mathcal{F}_Z := \left\{I \ : \ \|Z_I\| \leq \delta \cdot \sqrt{2\log(nK\rho/\delta_\star)}\right\} \tag{33}$$

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

$$\mathcal{G} := \mathcal{F}_\Phi \cup \mathcal{F}_\Psi \cup \mathcal{F}_Z. \tag{34}$$

We also need to control the sparse approximation step in each iteration. Recall that thresholding amounts to finding the largest S entries in magnitude of Ψ^*y, collecting them in the index set $\hat{I}$ and calculating the corresponding optimal coefficients as $\hat{x}_{\hat{I}} = \Psi_{\hat{I}}^\dagger y$. For the remainder of this chapter, we write

$$\mathcal{H} := \left\{(I, \sigma, c) \mid \hat{I} = I\right\} \tag{35}$$

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 can show that under the assumptions of Theorem 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 afterwards.

Lemma 4 *Under the assumptions of Theorem 3 we have*

$$[2\mathbb{P}(\mathcal{H}^c) + \mathbb{P}(\mathcal{G}^c)] \cdot \rho \leq \delta_\star/32. \tag{36}$$

Proof We begin with bounding the failure probability of thresholding, $\mathbb{P}(\mathcal{H}^c)$. Set $N := \Psi^*\Phi - \text{diag}(\Psi^*\Phi)$. 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 that $\underline{\alpha} = \min_i |\langle\psi_i, \phi_i\rangle|$. Plugging in the definition of y, we bound the failure probability as

$$\begin{aligned}
\mathbb{P}_y(\|\Psi_I^*y\|_{\min} < \|\Psi_{I^c}^*y\|_\infty) &= \mathbb{P}_y\left(\|\Psi_I^*\Phi_I x_I\|_{\min} < \|\Psi_{I^c}^*\Phi_I x_I\|_\infty\right) \\
&\leq \mathbb{P}_y\left(c_{\min}\|\text{diag}(\Psi_I^*\Phi_I)\|_{\min} - \|N_{I,I}x_I\|_\infty < \|\Psi_{I^c}^*\Phi_I x_I\|_\infty\right) \\
&\leq \mathbb{P}_y\left(c_{\min} \cdot \underline{\alpha} < 2\|N_I x_I\|_\infty\right) \\
&\leq \mathbb{P}_y\left(2\|N_I x_I\|_\infty \geq c_{\min} \cdot \underline{\alpha} \mid \|N_I\|_{\infty,2} < \eta\right) + \mathbb{P}_S\left(\|N_I\|_{\infty,2} \geq \eta\right). \tag{37}
\end{aligned}$$

15

To bound the first term, 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. As the signs σ are independent of the support I, we can apply Hoeffding's inequality to each entry of $N_I x_I$. The second term is a little more involved. By the Poissonisation trick [31, Lemma 3.5] we have

$$\mathbb{P}\left(\|N_I\|_{\infty,2} \geq \eta\right) \leq 2\mathbb{P}_B\left(\|N_I\|_{\infty,2} \geq \eta\right),$$

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, [42], restated in 11, together with an application of Hoeffding's inequality to the first term in (37) yields

$$\mathbb{P}_y(\|\Psi_I^* y\|_{\min} < \|\Psi_{I^c}^* y\|_\infty) \leq 2K \exp\left(-\frac{c_{\min}^2}{8c_{\max}^2 \eta^2} \cdot \underline{\alpha}^2\right) + 2K \left(e \frac{\|ND_{\sqrt{p}}\|_{\infty,2}^2}{\eta^2}\right)^{\frac{\eta^2}{\mu(\Psi,\Phi)^2}}$$

$$\leq 2K \exp\left(-\frac{\gamma^2}{8\eta^2} \cdot \underline{\alpha}^2\right) + 2K \left(2e \frac{\|ND_{\sqrt{\pi}}\|_{\infty,2}^2}{\eta^2}\right)^{\frac{\eta^2}{\mu(\Psi,\Phi)^2}}, \quad (38)$$

where we used that $(1 - \|p\|_\infty) \cdot p_i \leq \pi_i$, from [32] resp. Theorem 15(a), which implies $p_i \leq \frac{6}{5}\pi_i < 2\pi_i$ as well as $\|ND_{\sqrt{p}}\|_{\infty,2}^2 \leq 2\|ND_{\sqrt{\pi}}\|_{\infty,2}^2$.
Before going on we recall the abbreviations

$$\nu = \frac{1}{\sqrt{\log(nK\rho/\delta_\star)}} \leq \frac{1}{3} \quad \text{and} \quad \delta_\circ = \frac{\gamma}{C\log(nK\rho/\delta_\star)} = \frac{\gamma\nu^2}{C},$$

where the bound on ν holds true since $n\rho\delta_\star^{-1} \geq 130 \cdot 2 \cdot 42$. Setting $\eta := \gamma\underline{\alpha}\nu/4$ the first term in 38 becomes $2K\delta_\star^2/(n\rho K)^2$. To bound the second term observe that

$$\|ND_{\sqrt{\pi}}\|_{\infty,2}^2 = \|[\Psi^*\Phi - \text{diag}(\Psi^*\Phi)]D_{\sqrt{\pi}}\|_{\infty,2}^2 \leq \|\Phi D_{\sqrt{\pi}}\|^2, \quad (39)$$

so our condition on the generating dictionary in (13), $\|\Phi D_{\sqrt{\pi}}\| \leq \gamma\underline{\alpha}\nu/(4C)$, ensures that $2e\|ND_{\sqrt{\pi}}\|_{\infty,2}^2\eta^{-2} \leq e^{-2}$. To lower bound the exponent of the second term we will have to look at both regimes described in Theorem 3, separately. In the first regime, $\delta > \delta_\circ$, by (15) we simply have $\mu(\Psi,\Phi) \leq \underline{\alpha}\gamma\nu^2/(4C)$. In the second regime, $\delta \leq \delta_\circ$, we can employ the following bound

$$\mu(\Psi,\Phi) = \max_{i\neq j}|\langle\phi_i,\psi_j\rangle| \leq \max_{i\neq j}|\langle\phi_i,\phi_j\rangle| + \max_j\|\psi_j - \phi_j\| \leq \mu(\Phi) + \delta,$$

together with the observation that whenever $\delta \leq \delta_\circ \leq 1/C$, we have $\underline{\alpha} = 1 - \varepsilon^2/2 \geq 1 - \delta^2/2 \geq (C-1)/C$ and therefore also

$$\delta \leq \frac{\underline{\alpha}\gamma}{\underline{\alpha}C\log(nK\rho/\delta_\star)} \leq \frac{\underline{\alpha}\gamma}{(C-1)\log(nK\rho/\delta_\star)} = \frac{\underline{\alpha}\gamma\nu^2}{(C-1)}. \quad (40)$$

This means that in both regimes we have $\mu(\Psi,\Phi) \leq 2\underline{\alpha}\gamma\nu^2/(C-1)$. Substituting these bounds into 38, we get

$$\mathbb{P}\left(\mathcal{H}^c\right) \leq 2K\left(\frac{\delta_\star}{n\rho K}\right)^2 + 2K\left(\frac{\delta_\star}{n\rho K}\right)^{(C-1)^2/32} \leq 4K\left(\frac{\delta_\star}{n\rho K}\right)^2. \quad (41)$$

16

Now we turn to bounding the quantity $\mathbb{P}(\mathcal{F}_Z^c)$. Again by the Poissonisation trick, the matrix Chernoff inequality and the bound $p_i \leq 2\pi_i$ we have

$$\mathbb{P}\big(\|Z_I Z_I^*\| > t\big) \leq 2\mathbb{P}_B\big(\|Z_I Z_I^*\| > t\big)$$
$$\leq 2K\left(\frac{e\|ZD_p Z^*\|}{t}\right)^{t/\varepsilon^2} \leq 2K\left(\frac{2e\|ZD_\pi Z^*\|}{t}\right)^{t/\varepsilon^2} \leq 2K\left(\frac{2e\delta^2}{t}\right)^{t/\delta^2}.$$

Setting $t = 2\delta^2 \max\{e^2, \log(nK\rho/\delta_\star)\} = 2\delta^2 \log(nK\rho/\delta_\star)$ we get

$$\mathbb{P}(\mathcal{F}_Z^c) = \mathbb{P}\left(\|Z_I\| \geq \delta \cdot \sqrt{2\log(nK\rho/\delta_\star)}\right) \leq 2K\left(\frac{\delta_\star}{nK\rho}\right)^2. \tag{42}$$

Next we use Theorem 13, [31], and $p_i \leq \frac{6}{5}\pi_i$ to bound $\mathbb{P}(\mathcal{F}_\Phi^c \cup \mathcal{F}_\Psi^c)$ as

$$\mathbb{P}(\mathcal{F}_\Phi^c \cup \mathcal{F}_\Psi^c) \leq 512K \exp\left(-\min\left\{\frac{5\vartheta^2}{24e^2\|\Phi D_{\sqrt{\pi}}\|^2}, \frac{\vartheta}{2\mu(\Phi)}, \frac{5\vartheta^2}{24e^2\|\Psi D_{\sqrt{\pi}}\|^2}, \frac{\vartheta}{2\mu(\Psi)}\right\}\right).$$

In the first regime we have by assumption that $\max\{\nu\|\Phi D_{\sqrt{\pi}}\|, \mu(\Phi)\} \leq \gamma\underline{\alpha}\nu^2/(4C)$ and $\max\{\nu\|\Psi D_{\sqrt{\pi}}\|, \mu(\Psi)\} \leq \gamma\underline{\alpha}\nu^2/(4C)$, while in the second regime we only have the first inequality as well as $\delta \leq \gamma\underline{\alpha}\nu^2/(C-1)$. However, we can use the bound

$$\|\Psi D_{\sqrt{\pi}}\| \leq \|\Phi D_{\sqrt{\pi}}\| + \|(\Psi - \Phi)D_{\sqrt{\pi}}\| \leq \|\Phi D_{\sqrt{\pi}}\| + \delta$$
$$\leq \gamma\underline{\alpha}\nu \cdot \left(\tfrac{1}{4C} + \tfrac{\nu}{C-1}\right) \leq \gamma\underline{\alpha}\nu/C, \tag{43}$$

and the fact that for $\delta \leq \delta_\circ = \gamma\nu^2/C \leq \nu^2/C$ we have

$$\mu(\Psi) = \max_{i \neq j}|\langle \psi_i, \psi_j \rangle| \leq \max_{i \neq j}|\langle \phi_i, \phi_j \rangle + \langle \phi_i, \psi_j - \phi_j \rangle + \langle \psi_i - \phi_i, \psi_j \rangle|$$
$$\leq \mu(\Phi) + 2\varepsilon \leq \mu(\Phi) + 2\delta \leq \tfrac{9}{4} \cdot \nu^2/C. \tag{44}$$

All together this means that for $\vartheta = 1/4$ we can lower bound the expression in the minimum in both regimes by $2\log(nK\rho/\delta_\star)$, leading to $\mathbb{P}(\mathcal{F}_\Phi^c \cup \mathcal{F}_\Psi^c) \leq 512K\left(\delta_\star/(n\rho K)\right)^2$. Collecting all the bounds we finally get for $n \geq 130$ that

$$\mathbb{P}(\mathcal{H}^c) \cdot 2\rho + \mathbb{P}(\mathcal{G}^c) \cdot \rho \leq 522K\rho\left(\frac{\delta_\star}{n\rho K}\right)^2 \leq \frac{\delta_\star}{32}.$$

■

Another ingredient we need to prove the 4 inequalities used in the proof of our main theorem is the following lemma to estimate expectations of products of random matrices. A detailed proof, based on [9, 16], can be found in Appendix B.1 .

Lemma 5 *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* $\mathcal{I}$ *and* $\mathcal{G} \subseteq \mathcal{I}$. *If for all* $I \in \mathcal{G}$ *we have* $\|B(I)\| \leq \Gamma$ *then*

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

Finally, we also need the following corollary of results from [32], which is again proved in Appendix B.2.

Corollary 6 *Denote by $\mathbb{E}$ 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. Let $И$ be a $K \times K$ matrix with zero diagonal, $W = (w_1 \ldots, w_K)$ and $V = (v_1, \ldots, v_K)$ a pair of $d \times K$ matrices and $\mathcal{G}$ a subset of all supports of size S, meaning $\mathcal{G} \subseteq \{I : |I| = S\}$. If $\|\pi\|_\infty \leq 1/3$, we have*

$$\|\mathbb{E}[D_{\sqrt{\pi}}^{-1} R_I^* И_{I,I} R_I D_{\sqrt{\pi}}^{-1}]\| \leq 3 \cdot \|D_{\sqrt{\pi}} И D_{\sqrt{\pi}}\|, \tag{a}$$

$$\|\mathbb{E}[D_{\sqrt{\pi}}^{-1} R_I^* И_{I,I} И_{I,I}^* R_I D_{\sqrt{\pi}}^{-1}]\| \leq \tfrac{9}{2} \cdot \|D_{\sqrt{\pi}} И D_{\sqrt{\pi}}\|^2 + \tfrac{3}{2} \cdot \max_k \|e_k^* И D_{\sqrt{\pi}}\|^2, \tag{b}$$

$$\|\mathbb{E}[W R_I^* R_I V^* \cdot \mathbb{1}_I(\ell) \mathbb{1}_{\mathcal{G}}(I)]\| \leq \pi_\ell \cdot (\|W D_{\sqrt{\pi}}\| \cdot \|V D_{\sqrt{\pi}}\| + \|w_\ell\| \cdot \|v_\ell\|), \tag{c}$$

as well as

$$\|\mathbb{E}[D_{\sqrt{\pi}}^{-1} \mathbb{I}_{\ell^c} R_I^* И_{I,I} И_{I,I}^* R_I \mathbb{I}_{\ell^c} D_{\sqrt{\pi}}^{-1} \cdot \mathbb{1}_I(\ell)]\|$$
$$\leq \tfrac{3}{2} \cdot \pi_\ell \cdot \left(3 \cdot \|D_{\sqrt{\pi}} И e_\ell\|^2 + \max_k И_{k\ell}^2 + \tfrac{9}{2} \cdot \|D_{\sqrt{\pi}} И D_{\sqrt{\pi}}\|^2 + \tfrac{3}{2} \cdot \max_k \|e_k^* И D_{\sqrt{\pi}}\|^2 \right). \tag{d}$$

With the last three results in place we are finally able to prove Lemmas 7-10, which provide the 4 inequalities our proof is based on.

Lemma 7 *Under the assumptions of Theorem 3 we have*

$$\mathbb{P}\left(\|\Phi A(D_{\sqrt{\pi} \cdot \alpha \cdot \beta})^{-1} - \Phi D_{\sqrt{\pi}}\| > \underline{\alpha} \Delta/8 \right) \leq (d + K) \exp\left(-\frac{N(\Delta/16)^2}{2\rho^2 + \rho\Delta/16} \right). \tag{45}$$

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 the matrix Bernstein inequality to show that we 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 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

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

so we can use the matrix Bernstein inequality 12 from [42] to bound the left hand side of (45). For that we have to find an upper bound for the operator norm. By the assumptions on the generating dictionary and since we ensured in the algorithm that the estimated coefficients, which are too large are set to zero, meaning $\|\hat{x}_n\| \leq \kappa \|y_n\|$ and $\|y_n\| \leq S c_{\max}$, we get for $\rho = 2\kappa^2 S^2 \gamma^{-2} \underline{\alpha}^{-2} \underline{\pi}^{-3/2}$ and $\underline{\pi} < 1/3$,

$$\|\hat{Y}_n\| \leq \kappa S^2 c_{\max}^2 \|D_\beta^{-1}\| \|D_\alpha^{-1}\| \|D_{\sqrt{\pi}}^{-1}\| + \|\Phi D_{\sqrt{\pi}}\| \leq \rho \underline{\alpha}\underline{\pi} + 1 \leq \rho \underline{\alpha}/2 =: r. \tag{46}$$

18

Bounding $\|\mathbb{E}[\hat{Y}_n]\|$ 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. 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 $\Phi \operatorname{diag}(\mathbf{1}_{I_n}) D_{\sqrt{\pi}}^{-1}$

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

Note that $\|\Phi \operatorname{diag}(\mathbf{1}_{I_n}) D_{\sqrt{\pi}}^{-1}\| \le S \underline{\pi}^{-1/2}$ so the same bound as for $\|\hat{Y}_n\|$ holds. Concretely, 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 Y_n and $\hat{Y}_n$ coincide on $\mathcal{H}$, while the second terms coincide in expectation, meaning $\mathbb{E}[\Phi \operatorname{diag}(\mathbf{1}_{I_n}) D_{\sqrt{\pi}}^{-1}] = \Phi D_{\sqrt{\pi}}$. So dropping the index n for convenience, as each signal has the same distribution, e.g., writing I for I_n, we get using Lemma 4,

$$\|\mathbb{E}[\hat{Y}]\| \le \|\mathbb{E}[\hat{Y} - Y]\| + \|\mathbb{E}[Y]\| \le \mathbb{P}(\mathcal{H}^c) \cdot 2r + \|\mathbb{E}[\mathbb{1}_{\mathcal{G}^c}(I)Y]\| + \|\mathbb{E}[\mathbb{1}_{\mathcal{G}}(I)Y]\|$$
$$\le \mathbb{P}(\mathcal{H}^c) \cdot 2r + \mathbb{P}(\mathcal{G}^c) \cdot r + \|\mathbb{E}[\mathbb{1}_{\mathcal{G}}(I)Y]\|. \tag{47}$$

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

$$\|\mathbb{E}[\mathbb{1}_{\mathcal{G}}(I)Y]\| = \|\mathbb{E}_I[\mathbb{1}_{\mathcal{G}}(I) \cdot \mathbb{E}_{\sigma,c}[\Phi_I x_I x_I^* \Phi_I^*(\Psi_I^\dagger)^* R_I(D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1} - \Phi \operatorname{diag}(\mathbf{1}_I) D_{\sqrt{\pi}}^{-1}]]\|$$
$$= \|\mathbb{E}_{\mathcal{G}}[\Phi_I \Phi_I^*(\Psi_I^\dagger)^* R_I(D_{\sqrt{\pi}\cdot\alpha})^{-1} - \Phi_I(D_\alpha)_{I,I} R_I(D_{\sqrt{\pi}\cdot\alpha})^{-1}]\|. \tag{48}$$

We next have a closer look at the term $\Phi_I^*(\Psi_I^\dagger)$. Set $H = \mathbb{I} - \Psi^*\Psi$. For $I \in \mathcal{G}$ we have $\|H_{I,I}\| = \|\mathbb{I}_S - \Psi_I^*\Psi_I\| \le \vartheta$, meaning Ψ_I has full rank and $\Psi_I^\dagger = (\Psi_I^*\Psi_I)^{-1}\Psi_I^*$. We can further use the Neumann series to get the useful identity

$$(\Psi_I^*\Psi_I)^{-1} = (\mathbb{I}_S - H_{I,I})^{-1} = \sum_{k\ge 0} H_{I,I}^k = \mathbb{I}_S + (\Psi_I^*\Psi_I)^{-1}H_{I,I} \tag{49}$$
$$= \mathbb{I}_S + H_{I,I}(\Psi_I^*\Psi_I)^{-1}, \tag{50}$$

and the norm estimate $\|(\Psi_I^*\Psi_I)^{-1}\| \le (1-\vartheta)^{-1}$. By definition of $Z = \Psi - \Phi$, we have $\mathcal{E} := \operatorname{diag}(Z^*\Psi) = \mathbb{I} - D_\alpha$. We also define the zero diagonal matrix

$$\mathit{H} := (\Psi\mathcal{E} - Z)^*\Psi = (\Phi\mathcal{E} - ZD_\alpha)^*\Psi = (\Phi - \Psi D_\alpha)^*\Psi, \tag{51}$$

which lets us express $\Phi_I^*(\Psi_I^\dagger)^*$ as

$$\Phi_I^*(\Psi_I^\dagger)^* = (\Psi_I^* - Z_I^*)(\Psi_I^\dagger)^* = \mathbb{I}_S - Z_I^*(\Psi_I^\dagger)^*$$
$$= (D_\alpha)_{I,I} + \mathcal{E}_{I,I}\Psi_I^*\Psi_I(\Psi_I^*\Psi_I)^{-1} - Z_I^*\Psi_I(\Psi_I^*\Psi_I)^{-1}$$
$$= (D_\alpha)_{I,I} + \mathit{H}_{I,I}(\Psi_I^*\Psi_I)^{-1} = (D_\alpha)_{I,I} + \mathit{H}_{I,I} + \mathit{H}_{I,I}(\Psi_I^*\Psi_I)^{-1}H_{I,I}. \tag{52}$$

Note that for any I we have $\|\mathcal{H}_{I,I}\| \leq \|(\Psi\mathcal{E} - Z)_I\| \cdot \|\Psi_I\| \leq \varepsilon\sqrt{S} \cdot \sqrt{S} < 2S$ and therefore $\|\Phi_I \mathcal{H}_{I,I} R_I (D_{\sqrt{\pi}\cdot\alpha})^{-1}\| \leq \rho\underline{\alpha}/2 = r$. So substituting the expression for $\Phi_I^*(\Psi_I^\dagger)^*$ above into (48) resp. (47) and rewriting $\Phi_I = \Phi D_{\sqrt{\pi}} D_{\sqrt{\pi}}^{-1} R_I^*$ we get

$$\|\mathbb{E}[\hat{Y}]\| \leq [2\mathbb{P}(\mathcal{H}^c) + \mathbb{P}(\mathcal{G}^c)] \cdot r + \|\mathbb{E}_{\mathcal{G}}[\Phi_I(\mathcal{H}_{I,I} + \mathcal{H}_{I,I}(\Psi_I^*\Psi_I)^{-1} H_{I,I}) R_I (D_{\sqrt{\pi}\cdot\alpha})^{-1}]\|$$

$$\leq [2\mathbb{P}(\mathcal{H}^c) + \mathbb{P}(\mathcal{G}^c)] \cdot r + \mathbb{P}(\mathcal{G}^c) \cdot r + \|\mathbb{E}[\Phi_I \mathcal{H}_{I,I} R_I (D_{\sqrt{\pi}\cdot\alpha})^{-1}]\|$$

$$+ \|\mathbb{E}_{\mathcal{G}}[\Phi_I \mathcal{H}_{I,I}(\Psi_I^*\Psi_I)^{-1} H_{I,I} R_I (D_{\sqrt{\pi}\cdot\alpha})^{-1}]\|$$

$$\leq [\mathbb{P}(\mathcal{H}^c) + \mathbb{P}(\mathcal{G}^c)] \cdot \rho\underline{\alpha} + \|\Phi D_{\sqrt{\pi}}\| \cdot \|\mathbb{E}[D_{\sqrt{\pi}}^{-1} R_I^* \mathcal{H}_{I,I} R_I D_{\sqrt{\pi}}^{-1}]\| \cdot \|D_\alpha^{-1}\|$$

$$+ \|\Phi D_{\sqrt{\pi}}\| \cdot \|\mathbb{E}_{\mathcal{G}}[D_{\sqrt{\pi}}^{-1} R_I^* \mathcal{H}_{I,I}(\Psi_I^*\Psi_I)^{-1} H_{I,I} R_I D_{\sqrt{\pi}}^{-1}]\| \cdot \|D_\alpha^{-1}\|. \quad (53)$$

Using Corollary 6(a) and $\varepsilon \leq \min\{\delta, \sqrt{2}\}$, we bound the first expectation as

$$\|\mathbb{E}[D_{\sqrt{\pi}}^{-1} R_I^* \mathcal{H}_{I,I} R_I D_{\sqrt{\pi}}^{-1}]\| \leq 3\|D_{\sqrt{\pi}}\mathcal{H}D_{\sqrt{\pi}}\| \leq 3 \cdot (\|\Phi D_{\sqrt{\pi}}\| \cdot \varepsilon^2/2 + \|Z D_{\sqrt{\pi}}\|) \cdot \|\Psi D_{\sqrt{\pi}}\|$$

$$\leq 3 \cdot \delta \cdot \|\Psi D_{\sqrt{\pi}}\| \cdot (\|\Phi D_{\sqrt{\pi}}\| + 1). \quad (54)$$

Before we estimate the second expectation, note that applying Corollary 6(b) to $H, \mathcal{H}, \mathcal{H}^*$ and using that $\max\{\|\Psi D_{\sqrt{\pi}}\|, \|\Phi D_{\sqrt{\pi}}\|\} \leq \nu/C \leq 1/8$ yields the following three bounds, whose derivation can be found in Appendix B.2,

$$\|\mathbb{E}[D_{\sqrt{\pi}}^{-1} R_I^* H_{I,I} H_{I,I}^* R_I D_{\sqrt{\pi}}^{-1}]\| \leq 2 \cdot \|\Psi D_{\sqrt{\pi}}\|^2, \quad (55)$$

$$\|\mathbb{E}[D_{\sqrt{\pi}}^{-1} R_I^* \mathcal{H}_{I,I} \mathcal{H}_{I,I}^* R_I D_{\sqrt{\pi}}^{-1}]\| \leq 1/2 \cdot (3\|Z D_{\sqrt{\pi}}\| + 3\varepsilon)^2 \cdot \|\Psi D_{\sqrt{\pi}}\|^2, \quad (56)$$

$$\|\mathbb{E}[D_{\sqrt{\pi}}^{-1} R_I^* \mathcal{H}_{I,I}^* \mathcal{H}_{I,I} R_I D_{\sqrt{\pi}}^{-1}]\| \leq 2 \cdot (\|\Phi D_{\sqrt{\pi}}\| \cdot \varepsilon^2/2 + \|Z D_{\sqrt{\pi}}\|)^2. \quad (57)$$

Applying Theorem 5 to the second expectation in (53), using that on $\mathcal{G}$ we have $\|(\Psi_I^*\Psi_I)^{-1}\| \leq (1 - \vartheta)^{-1} \leq 4/3$, and the first two inequalities above yields

$$\|\mathbb{E}_{\mathcal{G}}[D_{\sqrt{\pi}}^{-1} R_I^* \mathcal{H}_{I,I} \cdot (\Psi_I^*\Psi_I)^{-1} \cdot H_{I,I} R_I D_{\sqrt{\pi}}^{-1}]\|$$

$$\leq \|\mathbb{E}[D_{\sqrt{\pi}}^{-1} R_I^* \mathcal{H}_{I,I} \mathcal{H}_{I,I}^* R_I D_{\sqrt{\pi}}^{-1}]\|^{1/2} \cdot 4/3 \cdot \|\mathbb{E}[D_{\sqrt{\pi}}^{-1} R_I^* H_{I,I}^* H_{I,I} R_I D_{\sqrt{\pi}}^{-1}]\|^{1/2}$$

$$\leq (4\|Z D_{\sqrt{\pi}}\| + 4\varepsilon) \cdot \|\Psi D_{\sqrt{\pi}}\|^2 \leq 8 \cdot \delta \cdot \|\Psi D_{\sqrt{\pi}}\|^2. \quad (58)$$

Substituting (54), (58) and the probability bound from Lemma 4 into (53) leads to

$$\|\mathbb{E}[\hat{Y}]\| \leq \underline{\alpha}\delta_\star/32 + \delta/\underline{\alpha} \cdot \|\Phi D_{\sqrt{\pi}}\| \cdot \|\Psi D_{\sqrt{\pi}}\| \cdot (3 + 3\|\Phi D_{\sqrt{\pi}}\| + 8\|\Psi D_{\sqrt{\pi}}\|). \quad (59)$$

By the assumptions of Theorem 3 we have in both regimes $\|\Phi D_{\sqrt{\pi}}\| \leq \gamma\underline{\alpha}\nu/(4C)$ and $\|\Psi D_{\sqrt{\pi}}\| \cdot \delta \leq \gamma\underline{\alpha}\nu/(2C)$, while due to (43) we have $\|\Psi D_{\sqrt{\pi}}\| \leq \gamma\underline{\alpha}\nu/C$ again in both regimes. Bounding the quantity in (59) in two ways we get

$$\|\mathbb{E}[\hat{Y}]\| \leq \frac{\alpha\delta_\star}{32} + \frac{\alpha\gamma}{C} \cdot \min\left\{\frac{\gamma\nu^2}{C}, \frac{2\gamma\nu^2}{C} \cdot \delta\right\} \leq \frac{\alpha\delta_\star}{32} + \frac{\alpha\gamma}{C} \cdot \min\{\delta_\circ, \delta\} \leq \underline{\alpha} \cdot \frac{\Delta}{16}.$$

Finally, an application of the matrix Bernstein inequality 12 with $t = m = \underline{\alpha}\Delta/16$ and $r = \underline{\alpha}\rho/2$ and some simplifications yield the desired bound. ∎

The next lemma shows that the matrix $B = \sum_{n=1}^N \hat{x}_n \hat{x}_n^*$ essentially behaves like a diagonal matrix. In particular, after rescaling we have $(D_{\sqrt{\pi}\cdot\alpha})^{-1} B (D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1} \approx \mathbb{I}$.

20

Lemma 8 *Under the assumptions of Theorem 3 we have*

$$\mathbb{P}\left(\|(D_{\sqrt{\pi}\cdot\alpha})^{-1}B(D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1} - \mathbb{I}\| > \Delta/4\right) \leq 2K\exp\left(-\frac{N(\Delta/16)^2}{2\rho^2 + \rho\Delta/16}\right).$$

Proof We will follow the approach in last proof very closely, that is, we write the matrix $(D_{\sqrt{\pi}\cdot\alpha})^{-1}B(D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1} - \mathbb{I}$ as a scaled sum of independent random matrices $\hat{Y}_n$ and apply the matrix Bernstein inequality. Recalling that $\hat{I}_n$ denotes the set found by 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

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

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

Recall that $\rho = 2\kappa^2 S^2\gamma^{-2}\underline{\alpha}^{-2}\underline{\pi}^{-3/2}$, so both matrices can be bounded as

$$\max\{\|\hat{Y}_n\|, \|Y_n\|\} \leq \kappa^2 S^2 c_{\max}^2 \|D_\beta^{-1}\|\|D_\alpha^{-2}\|\|D_\pi^{-1}\| + \|D_\pi^{-1}\| \leq 3\rho/4 =: r. \tag{60}$$

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 $\mathbb{E}[\text{diag}(\mathbf{1}_{I_n})D_\pi^{-1}] = \mathbb{I}$. So with the same argument as in (47) and as usual dropping the index n for convenience, we get

$$\|\mathbb{E}[\hat{Y}]\| \leq 2\rho \cdot \mathbb{P}(\mathcal{H}^c) + \rho \cdot \mathbb{P}(\mathcal{G}^c) + \|\mathbb{E}[\mathbb{1}_{\mathcal{G}}(I)Y]\|. \tag{61}$$

Similarly as in (48) 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) yields

$$\|\mathbb{E}[\mathbb{1}_{\mathcal{G}}(I)Y]\| = \|\mathbb{E}_{\mathcal{G}}\mathbb{E}_{\sigma,c}[(D_{\sqrt{\pi}\cdot\alpha})^{-1}R_I^*\Psi_I^\dagger\Phi_I x_I x_I^*\Phi_I^*(\Psi_I^\dagger)^*R_I(D_{\sqrt{\pi}\cdot\alpha\cdot\beta})^{-1} - \text{diag}(\mathbf{1}_I)D_\pi^{-1}]\|$$

$$= \left\|\mathbb{E}_{\mathcal{G}}\left[(D_{\sqrt{\pi}\cdot\alpha})^{-1}R_I^*\left(\Psi_I^\dagger\Phi_I^*\Phi_I^*(\Psi_I^\dagger)^* - (D_\alpha)_{I,I}^2\right)R_I(D_{\sqrt{\pi}\cdot\alpha})^{-1}\right]\right\|. \tag{62}$$

Using the expression for $\Phi_I^*(\Psi_I^\dagger)^*$ from (52) we get

$$\Psi_I^\dagger\Phi_I\Phi_I^*(\Psi_I^\dagger)^* - (D_\alpha^2)_{I,I} = (D_\alpha)_{I,I}H_{I,I}(\Psi_I^*\Psi_I)^{-1} + (\Psi_I^*\Psi_I)^{-1}H_{I,I}^*(D_\alpha)_{I,I}$$
$$+ (\Psi_I^*\Psi_I)^{-1}H_{I,I}^*H_{I,I}(\Psi_I^*\Psi_I)^{-1}. \tag{63}$$

The first two terms on the right hand side are each other's transpose, so substituting the above into (62) yields

$$\|\mathbb{E}[\mathbb{1}_{\mathcal{G}}(I)Y]\| \leq 2 \cdot \left\|\mathbb{E}_{\mathcal{G}}\left[D_{\sqrt{\pi}}^{-1}R_I^*H_{I,I}(\Psi_I^*\Psi_I)^{-1}R_ID_{\sqrt{\pi}}^{-1}\right]\right\| \cdot \|D_\alpha^{-1}\|$$
$$+ \|D_\alpha^{-1}\| \cdot \left\|\mathbb{E}_{\mathcal{G}}\left[D_{\sqrt{\pi}}^{-1}R_I^*(\Psi_I^*\Psi_I)^{-1}H_{I,I}^*H_{I,I}(\Psi_I^*\Psi_I)^{-1}R_ID_{\sqrt{\pi}}^{-1}\right]\right\| \cdot \|D_\alpha^{-1}\|. \tag{64}$$

To estimate the first term we repeat the steps in (53), noting that for all I we have $\|D_{\sqrt{\pi}}^{-1}R_I^*H_{I,I}R_ID_{\sqrt{\pi}}^{-1}\| \leq \rho\underline{\alpha}^2/2$. Using (54) as well as (58) we get

$$\|\mathbb{E}_{\mathcal{G}}D_{\sqrt{\pi}}^{-1}R_I^*H_{I,I}(\Psi_I^*\Psi_I)^{-1}R_ID_{\sqrt{\pi}}^{-1}]\| \leq \mathbb{P}(\mathcal{G}^c) \cdot \rho\underline{\alpha}^2/2 + \|\mathbb{E}[D_{\sqrt{\pi}}^{-1}R_I^*H_{I,I}R_ID_{\sqrt{\pi}}^{-1}]\|$$
$$+ \|\mathbb{E}_{\mathcal{G}}[D_{\sqrt{\pi}}^{-1}R_I^*H_{I,I}(\Psi_I^*\Psi_I)^{-1}H_{I,I}R_ID_{\sqrt{\pi}}^{-1}]\|$$
$$\leq \mathbb{P}(\mathcal{G}^c) \cdot \rho\underline{\alpha}^2/2 + 3 \cdot \left(\|\Phi D_{\sqrt{\pi}}\| \cdot \delta^2/2 + \|ZD_{\sqrt{\pi}}\|\right) \cdot \|\Psi D_{\sqrt{\pi}}\|$$
$$+ 4 \cdot \left(\|ZD_{\sqrt{\pi}}\| + \delta\right) \cdot \|\Psi D_{\sqrt{\pi}}\|^2. \tag{65}$$

Before we estimate the second term note that for $I \in \mathcal{G}$ we can use (51) to bound $\|\mathcal{U}_{I,I}(\Psi_I^*\Psi_I)^{-1}\|$ in two different ways, either as

$$\|\mathcal{U}_{I,I}(\Psi_I^*\Psi_I)^{-1}\| = \|(\Phi - \Psi D_\alpha)_I^* \Psi_I (\Psi_I^*\Psi_I)^{-1}\|$$
$$= \|\Phi_I(\Psi_I^\dagger)^* + (D_\alpha)_{I,I}\| \leq \sqrt{\tfrac{1+\vartheta}{1-\vartheta}} + 1 \leq \sqrt{5/3} + 1 \leq 7/3, \qquad (66)$$

or recalling that on $\mathcal{G}$ we have $\|Z_I\|^2 = \|Z_I^* Z_I\| \leq 2\delta^2 \log(nK\rho/\delta_\star) = 2\delta^2/\nu^2$ as

$$\|\mathcal{U}_{I,I}(\Psi_I^*\Psi_I)^{-1}\| = \|(\Psi\mathcal{E} - Z)_I^* \Psi_I (\Psi_I^*\Psi_I)^{-1}\|$$
$$= \|\mathcal{E}_{I,I} - Z_I^*(\Psi_I^\dagger)^*\| \leq \varepsilon^2/2 + \delta/\nu \cdot \sqrt{2} \leq \sqrt{2}\delta/\nu \cdot (\nu/2 + 1) \leq 2\delta/\nu. \quad (67)$$

We next split the second term using (49), apply Theorem 5 with the abbreviation $\Gamma = \min\{7/3, 2\delta/\nu\}$ and use Corollary 6(b) or rather (57) and (55) to get

$$\begin{aligned}
&\left\|\mathbb{E}_{\mathcal{G}}\big[D_{\sqrt{\pi}}^{-1} R_I^* (\Psi_I^*\Psi_I)^{-1} \mathcal{U}_{I,I}^* \mathcal{U}_{I,I}(\Psi_I^*\Psi_I)^{-1} R_I D_{\sqrt{\pi}}^{-1}\big]\right\| \\
&\leq \left\|\mathbb{E}\big[D_{\sqrt{\pi}}^{-1} R_I^* \mathcal{U}_{I,I}^* \mathcal{U}_{I,I} R_I D_{\sqrt{\pi}}^{-1}\big]\right\| \\
&\quad + 2\left\|\mathbb{E}_{\mathcal{G}}\big[D_{\sqrt{\pi}}^{-1} R_I^* \mathcal{U}_{I,I}^* \cdot \mathcal{U}_{I,I}(\Psi_I^*\Psi_I)^{-1} \cdot H_{I,I} R_I D_{\sqrt{\pi}}^{-1}\big]\right\| \\
&\quad + \left\|\mathbb{E}_{\mathcal{G}}\big[D_{\sqrt{\pi}}^{-1} R_I^* H_{I,I} \cdot (\Psi_I^*\Psi_I)^{-1} \mathcal{U}_{I,I}^* \mathcal{U}_{I,I}(\Psi_I^*\Psi_I)^{-1} \cdot H_{I,I} R_I D_{\sqrt{\pi}}^{-1}\big]\right\| \\
&\leq \left\|\mathbb{E}\big[D_{\sqrt{\pi}}^{-1} R_I^* \mathcal{U}_{I,I}^* \mathcal{U}_{I,I} R_I D_{\sqrt{\pi}}^{-1}\big]\right\| \\
&\quad + 2\left\|\mathbb{E}\big[D_{\sqrt{\pi}}^{-1} R_I^* \mathcal{U}_{I,I}^* \mathcal{U}_{I,I} R_I D_{\sqrt{\pi}}^{-1}\big]\right\|^{1/2} \cdot \Gamma \cdot \left\|\mathbb{E}\big[D_{\sqrt{\pi}}^{-1} R_I^* H_{I,I} H_{I,I} R_I D_{\sqrt{\pi}}^{-1}\big]\right\|^{1/2} \\
&\quad + \left\|\mathbb{E}\big[D_{\sqrt{\pi}}^{-1} R_I^* H_{I,I} H_{I,I} R_I D_{\sqrt{\pi}}^{-1}\big]\right\| \cdot \Gamma^2 \\
&= \left(\left\|\mathbb{E}\big[D_{\sqrt{\pi}}^{-1} R_I^* \mathcal{U}_{I,I}^* \mathcal{U}_{I,I} R_I D_{\sqrt{\pi}}^{-1}\big]\right\|^{1/2} + \Gamma \left\|\mathbb{E}\big[D_{\sqrt{\pi}}^{-1} R_I^* H_{I,I} H_{I,I} R_I D_{\sqrt{\pi}}^{-1}\big]\right\|^{1/2}\right)^2 \\
&\leq 2\left(\|\Phi D_{\sqrt{\pi}}\| \cdot \delta^2/2 + \|Z D_{\sqrt{\pi}}\| + \min\{7/3, 2\delta/\nu\} \cdot \|\Psi D_{\sqrt{\pi}}\|\right)^2. \qquad (68)
\end{aligned}$$

Substituting (65) and (68) into (64) and this in turn into (61) yields

$$\begin{aligned}
\|\mathbb{E}[\hat{Y}]\| \leq{}& 2\rho \cdot [\mathbb{P}(\mathcal{H}^c) + \mathbb{P}(\mathcal{G}^c)] + 6/\underline{\alpha} \cdot \left(\|\Phi D_{\sqrt{\pi}}\| \cdot \delta^2/2 + \|Z D_{\sqrt{\pi}}\|\right) \cdot \|\Psi D_{\sqrt{\pi}}\| \\
&+ 8/\underline{\alpha} \cdot \left(\|Z D_{\sqrt{\pi}}\| + \delta\right) \cdot \|\Psi D_{\sqrt{\pi}}\|^2 \\
&+ 2/\underline{\alpha}^2 \cdot \left(\|\Phi D_{\sqrt{\pi}}\| \cdot \delta^2/2 + \|Z D_{\sqrt{\pi}}\| + \min\{7/3, 2\delta/\nu\} \cdot \|\Psi D_{\sqrt{\pi}}\|\right)^2. \qquad (69)
\end{aligned}$$

We next show that the expression above is bounded by $\Delta/6$. In both regimes we have by Lemma (4) that $2\rho \cdot [\mathbb{P}(\mathcal{H}^c) + \mathbb{P}(\mathcal{G}^c)] \leq \delta_\star/16$. In the first regime, if $\delta > \delta_\circ = \gamma\nu^2/C$ and therefore $\min\{\delta_\circ, \delta\} = \delta_\circ$, we use that $\delta \leq \sqrt{2}$, $\max\{\|\Psi D_{\sqrt{\pi}}\|, \|\Phi D_{\sqrt{\pi}}\|\} \leq \underline{\alpha}\gamma\nu/(4C)$ and thus $\|Z D_{\sqrt{\pi}}\| \leq \underline{\alpha}\gamma\nu/(2C)$ to get

$$\|\mathbb{E}[\hat{Y}]\| \leq \frac{\delta_\star}{16} + \frac{\gamma\nu^2}{C} \cdot \frac{\gamma}{16C}\big[18\underline{\alpha} + 12\underline{\alpha} + 2 \cdot (16/3)^2\big] \leq \frac{\delta_\star}{16} + \delta_\circ \cdot \frac{87\gamma}{16C} \leq \frac{\Delta}{6}.$$

Conversely in the second regime, when $\delta \leq \delta_\circ$ and therefore $\min\{\delta_\circ, \delta\} = \delta$, we use that $\|Z D_{\sqrt{\pi}}\| \leq \delta \leq \delta_\circ = \gamma\nu^2/C$ and $\|\Psi D_{\sqrt{\pi}}\| \leq \underline{\alpha}\gamma\nu/C$ as well as $\nu \leq 1/2$ and $\underline{\alpha} \geq 1 - \delta_\circ^2/2 \geq 17/18$ to get

$$\|\mathbb{E}[\hat{Y}]\| \leq \frac{\delta_\star}{16} + \delta \cdot \frac{\gamma}{2C}\left[6 + \frac{6\underline{\alpha}\gamma^2}{64C^2} + \frac{8\underline{\alpha}\gamma}{4C} + \left(\frac{\gamma^2}{64C^2} + \frac{1}{\underline{\alpha}} + \frac{2\gamma}{C}\right)^2\right] \leq \frac{\delta_\star}{16} + \delta \cdot \frac{4\gamma}{C} \leq \frac{\Delta}{6}.$$

As before we arrive at the desired bound after applying the matrix Bernstein inequality 12 for $t = \Delta/16$ with $r = \frac{3}{4}\rho$ and $m = \Delta/6$ and some simplifications. ∎

Now we turn to bounding individual columns of the random matrices treated in the last two lemmas.

Lemma 9 *Under the conditions of Theorem 3 we have*

$$\mathbb{P}\Big(\|\Phi A(D_{\pi\cdot\alpha\cdot\beta})^{-1}e_\ell - \phi_\ell\| > \Delta/8\Big) \leq 28\exp\left(-\frac{N(\Delta/16)^2}{2\rho^2 + \rho\Delta/16}\right).$$

Proof As in the matrix case 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's inequality. To this end we define for a fixed index ℓ the random vectors

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

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

Note that we can obtain $\hat{Y}_n, Y_n$ by multiplying the analogue matrices in the proof of Lemma 7 from the right by $D_{\sqrt{\pi}}^{-1}e_\ell$. Following the proof strategy of Lemma 7 with the necessary changes, we first bound the ℓ_2-norm of the random vectors $\hat{Y}_n, Y_n$ as

$$\max\{\|\hat{Y}_n\|, \|Y_n\|\} \leq \kappa S^2 c_{\max}^2 \|\|D_\pi^{-1}\|\|D_\beta^{-1}\|\|D_\alpha^{-1}\| + S\|D_\pi^{-1}\| \leq 3\rho/4 =: r.$$

Repeating the procedures in (47) we next get

$$\|\mathbb{E}[\hat{Y}]\| \leq [2\mathbb{P}(\mathcal{H}^c) + \mathbb{P}(\mathcal{G}^c)] \cdot \rho + \|\mathbb{E}[\mathbb{1}_{\mathcal{G}}(I)Y]\| \leq \delta_\star/32 + \|\mathbb{E}[\mathbb{1}_{\mathcal{G}}(I)Y]\|, \qquad (70)$$

while repeating the procedures in (48)/(61), and using the expression from (52) $\Phi_I^*(\Psi_I^\dagger)^* = (D_\alpha)_{I,I} + \text{И}_{I,I}(\Psi_I^*\Psi_I)^{-1} = (D_\alpha)_{I,I} + \text{И}_{I,I} + \text{И}_{I,I}(\Psi_I^*\Psi_I)^{-1}H_{I,I}$ yields

$$\begin{aligned}
\|\mathbb{E}[\mathbb{1}_{\mathcal{G}}(I)Y]\| &= \|\mathbb{E}_{\mathcal{G}}\big[\Phi R_I^*\big(\Phi_I^*\Psi_I^{\dagger*} - (D_\alpha)_{I,I}\big)R_I e_\ell\big]\|/(\alpha_\ell \pi_\ell) \\
&\leq \|\mathbb{E}_{\mathcal{G}}\big[\Phi R_I^* \text{И}_{I,I}(\Psi_I^*\Psi_I)^{-1}R_I e_\ell\big]\|/(\underline{\alpha}\pi_\ell) \\
&\leq \left(\|\mathbb{E}_{\mathcal{G}}\big[\Phi R_I^* \text{И}_{I,I}R_I e_\ell\big]\| + \|\mathbb{E}_{\mathcal{G}}\big[\Phi R_I^* \text{И}_{I,I}(\Psi_I^*\Psi_I)^{-1}H_{I,I}R_I e_\ell\big]\|\right)/(\underline{\alpha}\pi_\ell). \quad (71)
\end{aligned}$$

We next use the decomposition $\mathbb{I} = \mathbb{I}_{\ell^c} + e_\ell e_\ell^*$. Note that for any diagonal matrix D we have $\mathbb{I}_{\ell^c}De_\ell = 0$ and that for any matrix $V_{\ell\ell} = e_\ell V e_\ell$. Since $\text{И}_{I,I}$ has a zero diagonal and we have $\text{И}_{I,I} = R_I(\Psi\mathcal{E} - Z)^*\Psi R_I^* = R_I(\Phi\mathcal{E} - ZD_\alpha)^*\Psi R_I^*$, Corollary 6(c) yields for the first term

$$\begin{aligned}
\|\Phi\mathbb{E}_{\mathcal{G}}\big[R_I^* \text{И}_{I,I}R_I e_\ell\big]\| &= \|\mathbb{E}_{\mathcal{G}}\big[\Phi\mathbb{I}_{\ell^c}R_I^* R_I(\Phi\mathcal{E} - ZD_\alpha)^* \mathbb{1}_I(\ell)\big] \cdot \psi_\ell\| \\
&\leq \pi_\ell \cdot \|\Phi\mathbb{I}_{\ell^c}D_{\sqrt{\pi}}\| \cdot \|(\Phi\mathcal{E} - ZD_\alpha)D_{\sqrt{\pi}}\| \cdot \|\psi_\ell\| \\
&\leq \pi_\ell \cdot \|\Phi D_{\sqrt{\pi}}\| \cdot (\|\Phi D_{\sqrt{\pi}}\| \cdot \varepsilon^2/2 + \|ZD_{\sqrt{\pi}}\|). \quad (72)
\end{aligned}$$

Before we estimate the second term with the same split, note that applying Corollary 6(d) to $H, \text{И}, \text{И}^*$ and using that $\max\{\|\Psi D_{\sqrt{\pi}}\|, \|\Phi D_{\sqrt{\pi}}\|\} \leq \nu/C \leq 1/8$ yields the following three

23

bounds, derived in detail in Appendix B.2,

$$\|\mathbb{E}[D_{\sqrt{\pi}}^{-1}\mathbb{I}_{\ell^c}R_I^*H_{I,I}H_{I,I}^*R_I\mathbb{I}_{\ell^c}D_{\sqrt{\pi}}^{-1}\cdot\mathbb{1}_I(\ell)]\| \leq 9\cdot\pi_\ell\cdot\max\{\mu(\Psi),\|\Psi D_{\sqrt{\pi}}\|\}^2, \tag{73}$$

$$\|\mathbb{E}[D_{\sqrt{\pi}}^{-1}\mathbb{I}_{\ell^c}R_I^*\mathcal{H}_{I,I}\mathcal{H}_{I,I}^*R_I\mathbb{I}_{\ell^c}D_{\sqrt{\pi}}^{-1}\cdot\mathbb{1}_I(\ell)]\| \leq 9\cdot\pi_\ell\cdot\delta^2, \tag{74}$$

$$\|\mathbb{E}[D_{\sqrt{\pi}}^{-1}\mathbb{I}_{\ell^c}R_I^*\mathcal{H}_{I,I}^*\mathcal{H}_{I,I}R_I\mathbb{I}_{\ell^c}D_{\sqrt{\pi}}^{-1}\cdot\mathbb{1}_I(\ell)]\| \leq 9\cdot\pi_\ell\cdot\delta^2. \tag{75}$$

Further, recall that $H_{I,\ell}=R_I He_\ell$ and $\mathcal{H}_{\ell,I}=e_\ell^*\mathcal{H}R_I^*$. So applying Theorem 5 and using (74) as well as twice Corollary 6(c), we get

$$
\begin{aligned}
\big\|\mathbb{E}_{\mathcal{G}}&\big[\Phi R_I^*\mathcal{H}_{I,I}(\Psi_I^*\Psi_I)^{-1}H_{I,I}R_I e_\ell\big]\big\| \\
&\leq \|\phi_\ell\|\cdot\big\|\mathbb{E}_{\mathcal{G}}\big[\mathcal{H}_{\ell,I}\cdot(\Psi_I^*\Psi_I)^{-1}\cdot H_{I,\ell}\,\mathbb{1}_I(\ell)\big]\big\| \\
&\qquad + \|\Phi D_{\sqrt{\pi}}\|\cdot\big\|\mathbb{E}_{\mathcal{G}}\big[D_{\sqrt{\pi}}^{-1}\mathbb{I}_{\ell^c}\mathcal{H}_{I,I}\cdot(\Psi_I^*\Psi_I)^{-1}\cdot H_{I,\ell}\,\mathbb{1}_I(\ell)\big]\big\| \\
&\leq \Big(\|\mathbb{E}[\mathcal{H}_{\ell,I}\mathcal{H}_{\ell,I}^*\mathbb{1}_I(\ell)]\|^{1/2} + \|\Phi D_{\sqrt{\pi}}\|\cdot\|\mathbb{E}[D_{\sqrt{\pi}}^{-1}\mathbb{I}_{\ell^c}\mathcal{H}_{I,I}\mathcal{H}_{I,I}^*\mathbb{I}_{\ell^c}D_{\sqrt{\pi}}^{-1}\mathbb{1}_I(\ell)]\|^{1/2}\Big) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\cdot 4/3\cdot\|\mathbb{E}[H_{I,\ell}^*H_{I,\ell}\mathbb{1}_I(\ell)]\|^{1/2} \\
&\leq (\sqrt{\pi_\ell}\cdot\|e_\ell^*\mathcal{H}D_{\sqrt{\pi}}\| + \|\Phi D_{\sqrt{\pi}}\|\cdot\sqrt{\pi_\ell}\cdot 3\delta)\cdot 4/3\cdot\sqrt{\pi_\ell}\cdot\|e_\ell^*HD_{\sqrt{\pi}}\| \\
&\leq 4/3\cdot\pi_\ell\cdot\delta\cdot(\|\Psi D_{\sqrt{\pi}}\| + 3\|\Phi D_{\sqrt{\pi}}\|)\cdot\|\Psi D_{\sqrt{\pi}}\|. \tag{76}
\end{aligned}
$$

Plugging the last two bounds back into (71) and (70) yields

$$
\begin{aligned}
\|\mathbb{E}[\hat{Y}]\| \leq \delta_\star/32 &+ 1/\underline{\alpha}\cdot\|\Phi D_{\sqrt{\pi}}\|\cdot(\|\Phi D_{\sqrt{\pi}}\|\cdot\varepsilon^2/2 + \|ZD_{\sqrt{\pi}}\|) \\
&+ 4/3\cdot\delta/\underline{\alpha}\cdot\|\Psi D_{\sqrt{\pi}}\|\cdot(\|\Psi D_{\sqrt{\pi}}\| + 3\|\Phi D_{\sqrt{\pi}}\|).
\end{aligned}
$$

We next proceed as in the last proof. We always have $\|\Phi D_{\sqrt{\pi}}\|\leq\underline{\alpha}\gamma\nu/(4C)$. If $\delta>\delta_\circ$, thus $\min\{\delta,\delta_\circ\}=\delta_\circ$, we use that additionally we have $\|\Psi D_{\sqrt{\pi}}\|\leq\underline{\alpha}\gamma\nu/(4C)$ and $\|ZD_{\sqrt{\pi}}\|\leq\underline{\alpha}\gamma\nu/(2C)$ to get

$$\|\mathbb{E}[\hat{Y}]\| \leq \frac{\delta_\star}{32} + \frac{\gamma\nu^2}{C}\cdot\frac{\underline{\alpha}\gamma}{16C}\cdot(1+2+4/3\cdot\sqrt{2}\cdot 4) \leq \frac{\delta_\star}{32} + \delta_\circ\cdot\frac{11\underline{\alpha}\gamma}{16C} \leq \frac{\Delta}{16}.$$

Conversely, if $\delta\leq\delta_\circ$, thus $\min\{\delta,\delta_\circ\}=\delta$, we use that $\|ZD_{\sqrt{\pi}}\|\leq\delta$ together with $\|\Psi D_{\sqrt{\pi}}\|\leq\underline{\alpha}\gamma\nu/C$ and $\nu\leq 1/3$ to get

$$\|\mathbb{E}[\hat{Y}]\| \leq \frac{\delta_\star}{32} + \delta\cdot\frac{\gamma\nu}{4C}\cdot\left(\frac{\underline{\alpha}\gamma\nu}{4C}\cdot\delta/2 + 1 + 4/3\cdot\frac{7\underline{\alpha}\gamma\nu}{C}\right) \leq \frac{\delta_\star}{32} + \delta\cdot\frac{2\gamma\nu}{4C} \leq \frac{\Delta}{16}.$$

Finally an application of the vector Bernstein inequality 12 for $t=m=\Delta/16$ and $r=3\rho/4$ and some simplifications yield the desired bound. ∎

Now to the grand final, showing that also the fourth inequality used in the proof of the main theorem is satisfied with high probability.

Lemma 10 *Under the conditions of Theorem 3 for $\Lambda=\max\left\{\|\Psi D_{\sqrt{\pi}}\|,\frac{\alpha\gamma\nu}{4C}\right\}$*

$$\mathbb{P}\left(\Lambda\cdot\|\mathbb{I}_{\ell^c}\left(D_{\sqrt{\pi}\cdot\alpha}\right)^{-1}B(D_{\pi\cdot\alpha\cdot\beta})^{-1}e_\ell\| > \frac{3\Delta}{16}\right) \leq 28\exp\left(-\frac{N(\Delta/16)^2}{2\rho^2+\rho\Delta/16}\right).$$

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 8, this time by multiplying from the left by $\mathbb{I}_{\ell^c}$ and from the right 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

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

while for the expectation we get similar to (48) and (62)

$$\|\mathbb{E}[\hat{Y}]\| \le \delta_\star/32 + \Lambda \cdot \pi^{-1}_\ell \cdot \|D^{-2}_\alpha\| \cdot \big\|\mathbb{E}_\mathcal{G}\big[D^{-1}_{\sqrt{\pi}}\mathbb{I}_{\ell^c} R^*_I\big(\Psi^\dagger_I \Phi_I \Phi^*_I \Psi^{\dagger*}_I\big) R_I e_\ell\big]\big\|. \tag{77}$$

Using a decomposition as in (63) we get

$$\mathbb{I}_{\ell^c} R^*_I\big(\Psi^\dagger_I \Phi_I \Phi^*_I \Psi^{\dagger*}_I\big) R_I e_\ell = -\mathbb{I}_{\ell^c} R^*_I \mathcal{H}_{I,I}(\Psi^*_I \Psi_I)^{-1} R_I e_\ell - \mathbb{I}_{\ell^c} R^*_I (\Psi^*_I \Psi_I)^{-1} \mathcal{H}^*_{I,I} R_I e_\ell$$
$$+ \mathbb{I}_{\ell^c} R^*_I (\Psi^*_I \Psi_I)^{-1} \mathcal{H}^*_{I,I} \mathcal{H}_{I,I} (\Psi^*_I \Psi_I)^{-1} R_I e_\ell. \tag{78}$$

The expectation corresponding to the first term can be obtained by replacing Φ with $D^{-1}_{\sqrt{\pi}}\mathbb{I}_{\ell^c}$ in (71). Going through (72) and (76) with the same change yields

$$\|\mathbb{E}_\mathcal{G}[D^{-1}_{\sqrt{\pi}}\mathbb{I}_{\ell^c} R^*_I \mathcal{H}_{I,I}(\Psi^*_I \Psi_I)^{-1} R_I e_\ell]\|$$
$$\le \pi_\ell \cdot (\|\Phi D_{\sqrt{\pi}}\| \cdot \varepsilon^2/2 + \|Z D_{\sqrt{\pi}}\| + 4\delta \cdot \|\Psi D_{\sqrt{\pi}}\|). \tag{79}$$

To estimate the second term note that $\mathcal{H}^*_{I,I} R_I e_\ell = R_I \mathbb{I}_{\ell^c} \mathcal{H}^* e_\ell \mathbb{1}_I(\ell)$ since $\mathcal{H}$ has a zero diagonal. So using the identity (50), Theorem 5, Corollary 6(c) and (d) or rather (73) we get for the second term

$$\|\mathbb{E}_\mathcal{G}[D^{-1}_{\sqrt{\pi}}\mathbb{I}_{\ell^c} R^*_I (\Psi^*_I \Psi_I)^{-1} \mathcal{H}^*_{I,I} R_I\, e_\ell]\|$$
$$\le \|\mathbb{E}_\mathcal{G}[D^{-1}_{\sqrt{\pi}} \mathbb{I}_{\ell^c} R^*_I (\Psi^*_I \Psi_I)^{-1} R_I \mathbb{I}_{\ell^c} D^{-1}_{\sqrt{\pi}} \mathbb{1}_I(\ell)]\| \cdot \|D_{\sqrt{\pi}} \mathcal{H}^* e_\ell\|$$
$$\le \varepsilon \cdot \|\Psi D_{\sqrt{\pi}}\| \cdot \big(\|\mathbb{E}_\mathcal{G}[D^{-1}_{\sqrt{\pi}}\mathbb{I}_{\ell^c} R^*_I R_I \mathbb{I}_{\ell^c} D^{-1}_{\sqrt{\pi}} \mathbb{1}_I(\ell)]\|$$
$$+ \|\mathbb{E}_\mathcal{G}[D^{-1}_{\sqrt{\pi}}\mathbb{I}_{\ell^c} R^*_I H_{I,I} \cdot (\Psi^*_I \Psi_I)^{-1} \cdot \mathbb{I}_{\ell^c} R_I D^{-1}_{\sqrt{\pi}} \mathbb{1}_I(\ell)]\|\big)$$
$$\le \delta \cdot \|\Psi D_{\sqrt{\pi}}\| \cdot \big(\pi_\ell + \|\mathbb{E}[D^{-1}_{\sqrt{\pi}}\mathbb{I}_{\ell^c} R^*_I H_{I,I} H^*_{I,I} R_I \mathbb{I}_{\ell^c} D^{-1}_{\sqrt{\pi}} \mathbb{1}_I(\ell)]\|^{1/2} \cdot 4/3 \cdot \sqrt{\pi_\ell}\big)$$
$$\le \pi_\ell \cdot \delta \cdot \|\Psi D_{\sqrt{\pi}}\| \cdot \big(1 + 4 \cdot \max\{\mu(\Psi), \|\Psi D_{\sqrt{\pi}}\|\}\big). \tag{80}$$

As probably feared the third term in (78) requires further decomposition. Again we split the inverse $(\Psi^*_I \Psi_I)^{-1}$ into $(\Psi^*_I \Psi_I)^{-1} = \mathbb{I} + (\Psi^*_I \Psi_I)^{-1} H_{I,I} = \mathbb{I} + H_{I,I}(\Psi^*_I \Psi_I)^{-1}$. Recalling that for $I \in \mathcal{G}$ we have $\|\mathcal{H}_{I,I}(\Psi^*_I \Psi_I)^{-1}\| \le \min\{7/3, 2\delta/\nu\} = \Gamma$ and applying Theorem 5 to

25

all four resulting terms yields

$$\|\mathbb{E}_{\mathcal{G}}[D_{\sqrt{\pi}}^{-1}\mathbb{I}_{\ell^c}R_I^*\mathcal{H}_{I,I}^* \cdot \mathcal{H}_{I,I}R_I\,e_\ell]\|$$
$$\leq \|\mathbb{E}[D_{\sqrt{\pi}}^{-1}\mathbb{I}_{\ell^c}R_I^*\mathcal{H}_{I,I}^*\mathcal{H}_{I,I}R_I\mathbb{I}_{\ell^c}D_{\sqrt{\pi}}^{-1}\mathbb{1}_I(\ell)]\|^{1/2} \cdot \|\mathbb{E}[\mathcal{H}_{I,\ell}^*\mathcal{H}_{I,\ell}\mathbb{1}_I(\ell)]\|^{1/2}$$

$$\|\mathbb{E}_{\mathcal{G}}[D_{\sqrt{\pi}}^{-1}\mathbb{I}_{\ell^c}R_I^*H_{I,I} \cdot (\Psi_I^*\Psi_I)^{-1}\mathcal{H}_{I,I}^* \cdot \mathcal{H}_{I,I}R_I\,e_\ell]\|$$
$$\leq \|\mathbb{E}[D_{\sqrt{\pi}}^{-1}\mathbb{I}_{\ell^c}R_I^*H_{I,I}H_{I,I}^*R_I\mathbb{I}_{\ell^c}D_{\sqrt{\pi}}^{-1}\mathbb{1}_I(\ell)]\|^{1/2} \cdot \Gamma \cdot \|\mathbb{E}[\mathcal{H}_{I,\ell}^*\mathcal{H}_{I,\ell}\mathbb{1}_I(\ell)]\|^{1/2}$$

$$\|\mathbb{E}_{\mathcal{G}}[D_{\sqrt{\pi}}^{-1}\mathbb{I}_{\ell^c}R_I^*\mathcal{H}_{I,I}^* \cdot \mathcal{H}_{I,I}(\Psi_I^*\Psi_I)^{-1} \cdot H_{I,I}R_Ie_\ell]\|$$
$$\leq \|\mathbb{E}[D_{\sqrt{\pi}}^{-1}\mathbb{I}_{\ell^c}R_I^*\mathcal{H}_{I,I}^*\mathcal{H}_{I,I}R_I\mathbb{I}_{\ell^c}D_{\sqrt{\pi}}^{-1}\mathbb{1}_I(\ell)]\|^{1/2} \cdot \Gamma \cdot \|\mathbb{E}[H_{I,\ell}^*H_{I,\ell}\mathbb{1}_I(\ell)]\|^{1/2}$$

$$\|\mathbb{E}_{\mathcal{G}}[D_{\sqrt{\pi}}^{-1}\mathbb{I}_{\ell^c}R_I^*H_{I,I} \cdot (\Psi_I^*\Psi_I)^{-1}\mathcal{H}_{I,I}^*\mathcal{H}_{I,I}(\Psi_I^*\Psi_I)^{-1} \cdot H_{I,I}R_Ie_\ell]\|$$
$$\leq \|\mathbb{E}[D_{\sqrt{\pi}}^{-1}\mathbb{I}_{\ell^c}R_I^*H_{I,I}H_{I,I}^*R_I\mathbb{I}_{\ell^c}D_{\sqrt{\pi}}^{-1}\mathbb{1}_I(\ell)]\|^{1/2} \cdot \Gamma^2 \cdot \|\mathbb{E}[H_{I,\ell}^*H_{I,\ell}\mathbb{1}_I(\ell)]\|^{1/2}$$

Bounding the terms on the left hand side via Corollary 6(d) or rather (73) and (75) and the terms on the right hand side via Corollary 6(c) we get

$$\|\mathbb{E}_{\mathcal{G}}[D_{\sqrt{\pi}}^{-1}\mathbb{I}_{\ell^c}R_I^*(\Psi_I^*\Psi_I)^{-1}\mathcal{H}_{I,I}^*\mathcal{H}_{I,I}(\Psi_I^*\Psi_I)^{-1}R_Ie_\ell]\|$$
$$\leq \pi_\ell \cdot \big(3\delta + 3\Gamma\max\{\mu(\Psi),\|\Psi D_{\sqrt{\pi}}\|\}\big)\big(\|e_\ell^*\mathcal{H}D_{\sqrt{\pi}}\| + \Gamma\|e_\ell^*HD_{\sqrt{\pi}}\|\big)$$
$$\leq \pi_\ell \cdot \big(3\delta + 3\Gamma\max\{\mu(\Psi),\|\Psi D_{\sqrt{\pi}}\|\}\big)\big(\|\Phi D_{\sqrt{\pi}}\|\,\varepsilon^2/2 + \|Z D_{\sqrt{\pi}}\| + \Gamma\|\Psi D_{\sqrt{\pi}}\|\big). \quad (81)$$

Finally substituting (79)-(81) into (77) yields

$$\|\mathbb{E}[\hat{Y}]\| \leq \frac{\delta_\star}{32} + \frac{\Lambda}{\underline{\alpha}^2}\Big(\|\Phi D_{\sqrt{\pi}}\|\frac{\varepsilon^2}{2} + \|ZD_{\sqrt{\pi}}\| + \delta\,\|\Psi D_{\sqrt{\pi}}\|\,\big(5 + 4\max\{\mu(\Psi),\|\Psi D_{\sqrt{\pi}}\|\}\big)\Big)$$
$$+ \frac{3\Lambda}{\underline{\alpha}^2}\Big(\delta + \Gamma\max\{\mu(\Psi),\|\Psi D_{\sqrt{\pi}}\|\}\Big)\Big(\|\Phi D_{\sqrt{\pi}}\|\frac{\varepsilon^2}{2} + \|ZD_{\sqrt{\pi}}\| + \Gamma\|\Psi D_{\sqrt{\pi}}\|\Big).$$

As before we distinguish between $\delta > \delta_\circ$, where $\min\{\delta,\delta_\circ\} = \delta_\circ = \gamma\nu^2/C$ and we can use $\|\Psi D_{\sqrt{\pi}}\| \leq \underline{\alpha}\gamma\nu/(4C) = \Lambda$, $\mu(\Psi) \leq \underline{\alpha}\gamma\nu^2/(4C)$ as well as $\|ZD_{\sqrt{\pi}}\| \leq \underline{\alpha}\gamma\nu/(2C)$ and $\Gamma \leq 7/3$, to get

$$\|\mathbb{E}[\hat{Y}]\| \leq \frac{\delta_\star}{32} + \frac{\gamma^2\nu^2}{16C^2}\Big[3 + \sqrt{2}\Big(5 + \frac{\underline{\alpha}\gamma\nu}{C}\Big) + 16\Big(\sqrt{2} + \frac{7\underline{\alpha}\gamma\nu}{12C}\Big)\Big] \leq \frac{\delta_\star}{32} + \delta_\circ\frac{36\gamma}{16C} \leq \frac{\Delta}{16},$$

and $\delta \leq \delta_\circ$, where $\min\{\delta,\delta_\circ\} = \delta$ and we can use $\|\Psi D_{\sqrt{\pi}}\| \leq \underline{\alpha}\gamma\nu/C = \Lambda$ and $\mu(\Psi) \leq 9\nu^2/(4C)$, meaning $\max\{\mu(\Psi),\|\Psi D_{\sqrt{\pi}}\|\} \leq \nu \leq 1$ as well as $\|ZD_{\sqrt{\pi}}\| \leq \delta$ and $\Gamma \leq 2\delta/\nu$, to again get

$$\|\mathbb{E}[\hat{Y}]\| \leq \frac{\delta_\star}{32} + \delta\frac{\gamma\nu}{\underline{\alpha}C}\Big[1 + 9\delta + \frac{9\underline{\alpha}\gamma}{8C}\big(8\nu + \delta\nu + \delta^2\nu + 16\delta\big)\Big] \leq \frac{\delta_\star}{32} + \delta\frac{2\gamma\nu}{\underline{\alpha}C} \leq \frac{\Delta}{16}.$$

As before an application of the vector Bernstein inequality 12 for $t = m = \Delta/16$ and $r = 3\rho/4$ and some simplifications yield the desired bound. ∎

Note that for the probability bounds in the last four lemmas, we have used the simplified version of the vector/matrix Bernstein inequality from [42], stated in 12, which bounds the variance appearing in the denominator via the quantity r. Using the same approach as above one could estimate terms of the form $\|\tilde{Y}_n\tilde{Y}_n^*\|$ and $\|\tilde{Y}_n^*\tilde{Y}_n\|$, where $\tilde{Y}_n$ is the centered version of Y_n, to improve the bounds on the variance and thus lower the number of training signals required in the main theorem, [21, 42].

Appendix B. Technical results

For convenience we first restate several bounds for random matrices used in the proofs of Lemma 4 as well as Lemma 7-10. We then provide the proofs of Lemma 5 and Corollary 6 together with the detailed derivation of (55-57) and (73-75).

B.1 Random matrix bounds

We first recall two bounds for sums of random matrices and vectors.

Theorem 11 (Matrix Chernoff inequality [42]) *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 $t \geq e\mu_{max}$,*

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

Theorem 12 (Matrix resp. vector Bernstein inequality [42, 21]) *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 $t > 0$,

$$\mathbb{P}\left(\|\frac{1}{N}\sum_{n=1}^N Y_n\| \geq m + t\right) \leq \kappa \exp\left(\frac{-Nt^2}{2r^2 + (r+m)t}\right), \tag{82}$$

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

Next we recall Theorem 3.1 from [31] which allows us to control the operator norm of a submatrix with high probability.

Theorem 13 (Operator norm of a random submatrix [31]) *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 $\sum_{i=1}^K p_i = S$. Further let D_p denote the diagonal matrix with the vector p on its diagonal. Then*

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

To prove Theorem 5 we first state and prove the following lemma to bound sums of products of matrices.

Lemma 14 (Sums of products of matrices [9], [16]) *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

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

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

Lemma 5 is a straightforward consequence of the result above.

Proof [of Lemma 5] We want to show that for random matrices $A(I) \in \mathbb{R}^{d_1 \times d_2}$, $B(I) \in \mathbb{R}^{d_2 \times d_3}$, $C(I) \in \mathbb{R}^{d_3 \times d_4}$, where I is a discrete random variable taking values in $\mathcal{I}$ and $\mathcal{G} \subseteq \mathcal{I}$ with $\max_{I \in \mathcal{G}} \|B(I)\| \leq \Gamma$ we have

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

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

$$\|\mathbb{E}\left[A(I)B(I)C(I)\mathbb{1}_{\mathcal{G}}(I)\right]\| = \|\sum_{I \in \mathcal{G}} \mathbb{P}[I]^{1/2}A(I)B(I)C(I)\mathbb{P}[I]^{1/2}\|$$

$$\leq \|\sum_{I \in \mathcal{G}} \mathbb{P}[I]A(I)A(I)^*\|^{1/2} \cdot \Gamma \cdot \|\sum_{I \in \mathcal{G}} \mathbb{P}[I]C(I)^*C(I)\|^{1/2}$$

$$\leq \|\mathbb{E}\left[A(I)A(I)^*\right]\|^{1/2} \cdot \Gamma \cdot \|\mathbb{E}\left[C(I)^*C(I)\right]\|^{1/2},$$

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 $\mathbb{P}[I] \geq 0$. ∎

B.2 Proof of Corollary 6

Finally, we will turn to the proof of Corollary 6. Note that in the uniformly distributed support model, which is equivalent to rejective sampling with uniform weights $p_i = S/K$, most estimates become trivial, since we have $\pi_i = \mathbb{P}(i \in I) = S/K = p_i$ and $\mathbb{P}(\{i, j\} \subseteq I) = \frac{S(S-1)}{K(K-1)}$. This means that for a zero diagonal matrix M we have

$$\mathbb{E}[R_I^* M_{I,I} R_I] = \mathbb{E}[M \odot (\mathbf{1}_I \mathbf{1}_I^*)] = M \odot \mathbb{E}[\mathbf{1}_I \mathbf{1}_I^*] = M \cdot \frac{S(S-1)}{K(K-1)},$$

so we simply get

$$\|\mathbb{E}[D_{\sqrt{\pi}}^{-1} R_I^* M_{I,I} R_I D_{\sqrt{\pi}}^{-1}]\| = \|M\| \cdot \frac{S-1}{K-1} \leq \|M\| \cdot \frac{S}{K} = \|D_{\sqrt{\pi}} M D_{\sqrt{\pi}}\|.$$

Unfortunately, in the rejective sampling model with non-uniform weights p_i, these estimates become much more involved. In particular, we will heavily rely on the following theorem, collecting results from [32].

Theorem 15 *Let $\mathbb{P}_B$ be the probability measure corresponding to the Poisson sampling model 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 in Definition 1. Further denote by $\mathbb{E}_S$ the expectation with respect to $\mathbb{P}_S$ and 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*

$$(1 - \|p\|_\infty) \cdot p_i \leq \pi_S(i) \leq 2 \cdot p_i, \quad \text{if } \sum_k p_k = S, \tag{a}$$

$$\pi_{S-1}(i) \leq \pi_S(i), \tag{b}$$

$$\mathbb{P}_S(\{i,j\} \subseteq I) \leq \pi_S(i) \cdot \pi_S(j), \quad \text{if } i \neq j. \tag{c}$$

Further, defining for $L \subseteq [K]$ with $|L| < S$ the set $\mathcal{L} = \{I \subseteq [K] : L \subseteq I\}$, we have

$$\mathbb{E}_S\big[\mathbf{1}_{I \setminus L}\mathbf{1}_{I \setminus L}^* \cdot \mathbb{1}_{\mathcal{L}}(I)\big] \cdot \prod_{\ell \in L}[1 - \pi_S(\ell)] \preceq \mathbb{E}_{S-|L|}[\mathbf{1}_I \mathbf{1}_I^*] \cdot \prod_{\ell \in L} \pi_S(\ell). \tag{d}$$

Finally, if $\pi := \pi_S$ satisfies $\|\pi\|_\infty < 1$, then for any $K \times K$ matrix A we have

$$\|A \odot \mathbb{E}[\mathbf{1}_I \mathbf{1}_I^*]\| \leq \frac{1 + \|\pi\|_\infty}{(1 - \|\pi\|_\infty)^2} \cdot \|D_\pi[A - \operatorname{diag}(A)]D_\pi\| + \|\operatorname{diag}(A)D_\pi\|. \tag{e}$$

With these results in hand we can finally prove Corollary 6.

Proof [of Corollary 6 including (55-57) and (73-75)]

(a) We want to show that for a matrix $\mathcal{U}$ with zero diagonal we have

$$\|\mathbb{E}[D_{\sqrt{\pi}}^{-1} R_I^* \mathcal{U}_{I,I} R_I D_{\sqrt{\pi}}^{-1}]\| \leq 3 \cdot \|D_{\sqrt{\pi}} \mathcal{U} D_{\sqrt{\pi}}\|.$$

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

$$DR_I^* A_{I,I} R_I D = DR_I^* R_I A R_I^* R_I D = D \operatorname{diag}(\mathbf{1}_I) A \operatorname{diag}(\mathbf{1}_I) D$$
$$= \operatorname{diag}(\mathbf{1}_I) D A D \operatorname{diag}(\mathbf{1}_I) = (DAD) \odot (\mathbf{1}_I \mathbf{1}_I^*).$$

Using Theorem 15(e) and $\|\pi\|_\infty \leq 1/3$ we therefore get for $\mathcal{U}$ with zero-diagonal

$$\|\mathbb{E}[D_{\sqrt{\pi}}^{-1} R_I^* \mathcal{U}_{I,I} R_I D_{\sqrt{\pi}}^{-1}]\| = \|\mathbb{E}[(D_{\sqrt{\pi}}^{-1} \mathcal{U} D_{\sqrt{\pi}}^{-1}) \odot (\mathbf{1}_I \mathbf{1}_I^*)]\|$$
$$= \|(D_{\sqrt{\pi}}^{-1} \mathcal{U} D_{\sqrt{\pi}}^{-1}) \odot \mathbb{E}[\mathbf{1}_I \mathbf{1}_I^*]\|$$
$$\leq 3\|D_\pi D_{\sqrt{\pi}}^{-1} \mathcal{U} D_{\sqrt{\pi}}^{-1} D_\pi\| = 3\|D_{\sqrt{\pi}} \mathcal{U} D_{\sqrt{\pi}}\|,$$

which proves (a). **(a)**✓

(b) We want to show that for a matrix $\mathcal{U}$ with zero diagonal we have

$$\|\mathbb{E}[D_{\sqrt{\pi}}^{-1} R_I^* \mathcal{U}_{I,I} \mathcal{U}_{I,I}^* R_I D_{\sqrt{\pi}}^{-1}]\| \leq \tfrac{9}{2} \cdot \|D_{\sqrt{\pi}} \mathcal{U} D_{\sqrt{\pi}}\|^2 + \tfrac{3}{2} \cdot \max_k \|e_k^* \mathcal{U} D_{\sqrt{\pi}}\|^2.$$

We again rewrite the expression, whose expectation we need to estimate, as

$$R_I^* \mathcal{U}_{I,I} \mathcal{U}_{I,I}^* R_I = R_I^* R_I \cdot \mathcal{U} \cdot R_I^* R_I \cdot \mathcal{U}^* \cdot R_I^* R_I$$
$$= \operatorname{diag}(\mathbf{1}_I) \cdot \mathcal{U} \cdot \operatorname{diag}(\mathbf{1}_I) \cdot \mathcal{U}^* \cdot \operatorname{diag}(\mathbf{1}_I)$$
$$= [\mathcal{U} \cdot \operatorname{diag}(\mathbf{1}_I) \cdot \mathcal{U}^*] \odot (\mathbf{1}_I \mathbf{1}_I^*)$$
$$= \Big(\sum_{k \in I} \mathcal{U}_k \mathcal{U}_k^*\Big) \odot (\mathbf{1}_I \mathbf{1}_I^*) = \sum_{k \in I} (\mathcal{U}_k \mathcal{U}_k^*) \odot (\mathbf{1}_I \mathbf{1}_I^*).$$

29

Since the k-th entry of $\mathcal{H}_k$ and therefore both the k-th row and k-th column of $\mathcal{H}_k\mathcal{H}_k^*$ are zero, we have $(\mathcal{H}_k\mathcal{H}_k^*) \odot (\mathbf{1}_I\mathbf{1}_I^*) = (\mathcal{H}_k\mathcal{H}_k^*) \odot (\mathbf{1}_{I\backslash\{k\}}\mathbf{1}_{I\backslash\{k\}}^*)$, yielding

$$
\begin{aligned}
R_I^*\mathcal{H}_{I,I}\mathcal{H}_{I,I}^*R_I &= \sum_{k \in I}(\mathcal{H}_k\mathcal{H}_k^*) \odot (\mathbf{1}_{I\backslash\{k\}}\mathbf{1}_{I\backslash\{k\}}^*) \\
&= \sum_k (\mathbb{1}_I(k) \cdot \mathcal{H}_k\mathcal{H}_k^*) \odot (\mathbf{1}_{I\backslash\{k\}}\mathbf{1}_{I\backslash\{k\}}^*) \\
&= \sum_k (\mathcal{H}_k\mathcal{H}_k^*) \odot (\mathbf{1}_{I\backslash\{k\}}\mathbf{1}_{I\backslash\{k\}}^* \cdot \mathbb{1}_I(k)).
\end{aligned}
\tag{83}
$$

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 15(d) further leads to

$$
\begin{aligned}
\mathbb{E}_S\left[R_I^*\mathcal{H}_{I,I}\mathcal{H}_{I,I}^*R_I\right] &= \sum_k (\mathcal{H}_k\mathcal{H}_k^*) \odot \mathbb{E}_S\left[\mathbf{1}_{I\backslash\{k\}}\mathbf{1}_{I\backslash\{k\}}^* \cdot \mathbb{1}_I(k)\right] \\
&\preceq \sum_k (\mathcal{H}_k\tfrac{\pi_S(k)}{1-\pi_S(k)}\mathcal{H}_k^*) \odot \mathbb{E}_{S-1}[\mathbf{1}_I\mathbf{1}_I^*] \\
&= \Big(\sum_k \mathcal{H}_k\tfrac{\pi_S(k)}{1-\pi_S(k)}\mathcal{H}_k^*\Big) \odot \mathbb{E}_{S-1}[\mathbf{1}_I\mathbf{1}_I^*] \\
&= (\mathcal{H}\operatorname{diag}(\tfrac{\pi_S}{1-\pi_S})\mathcal{H}^*) \odot \mathbb{E}_{S-1}[\mathbf{1}_I\mathbf{1}_I^*].
\end{aligned}
$$

Abbreviating $M := \mathcal{H}\operatorname{diag}(\tfrac{\pi_S}{1-\pi_S})\mathcal{H}^*$, and using Theorem 15 (e) and (b) we get

$$
\begin{aligned}
\|\mathbb{E}_S\left[D_{\sqrt{\pi}}^{-1}R_I^*\mathcal{H}_{I,I}\mathcal{H}_{I,I}^*R_ID_{\sqrt{\pi}}^{-1}\right]\| &= \|D_{\sqrt{\pi}}^{-1}\mathbb{E}_S\left[R_I^*\mathcal{H}_{I,I}\mathcal{H}_{I,I}^*R_I\right]D_{\sqrt{\pi}}^{-1}\| \\
&\leq \|(D_{\sqrt{\pi}}^{-1}MD_{\sqrt{\pi}}^{-1}) \odot \mathbb{E}_{S-1}[\mathbf{1}_I\mathbf{1}_I^*]\| \\
&\leq 3\|D_{\pi_{S-1}}[D_{\sqrt{\pi}}^{-1}MD_{\sqrt{\pi}}^{-1} - \operatorname{diag}(D_{\sqrt{\pi}}^{-1}MD_{\sqrt{\pi}}^{-1})]D_{\pi_{S-1}}\| \\
&\qquad\qquad + \|\operatorname{diag}(D_{\sqrt{\pi}}^{-1}MD_{\sqrt{\pi}}^{-1})D_{\pi_{S-1}}\| \\
&\leq 3\|D_{\sqrt{\pi}}MD_{\sqrt{\pi}} - \operatorname{diag}(D_{\sqrt{\pi}}MD_{\sqrt{\pi}})\| + \|\operatorname{diag}(M)\| \\
&\leq 3\|D_{\sqrt{\pi}}MD_{\sqrt{\pi}}\| + \|\operatorname{diag}(M)\|,
\end{aligned}
$$

where in last inequality we have used that $D_{\sqrt{\pi}}MD_{\sqrt{\pi}}$ is positive semidefinite. Combining the inequality above with the bounds

$$
\begin{aligned}
\|D_{\sqrt{\pi}}MD_{\sqrt{\pi}}\| &= \|D_{\sqrt{\pi}}\mathcal{H}D_{\sqrt{\pi}}\operatorname{diag}(\tfrac{1}{1-\pi})D_{\sqrt{\pi}}\mathcal{H}^*D_{\sqrt{\pi}}\| \leq (1-\|\pi\|_\infty)^{-1}\|D_{\sqrt{\pi}}\mathcal{H}D_{\sqrt{\pi}}\|^2, \\
\|\operatorname{diag}(M)\| &= \max_k e_k^*\mathcal{H}D_{\sqrt{\pi}}\operatorname{diag}(\tfrac{1}{1-\pi})D_{\sqrt{\pi}}\mathcal{H}^*e_k \leq (1-\|\pi\|_\infty)^{-1}\max_k \|e_k^*\mathcal{H}D_{\sqrt{\pi}}\|^2,
\end{aligned}
$$

and our assumption that $\|\pi\|_\infty \leq 1/3$ leads to (b). $\qquad\qquad$ **(b)**✓

(55-57) We next specialise the general inequality from (b) to the three concrete matrices needed in the proofs of Lemma 7-10. For the case $H = \mathbb{I} - \Psi^*\Psi$, note that since $D_{\sqrt{\pi}}\Psi^*\Psi D_{\sqrt{\pi}}$ is a positive semidefinite matrix we have

$$
\|D_{\sqrt{\pi}}HD_{\sqrt{\pi}}\| = \|D_{\sqrt{\pi}}\Psi^*\Psi D_{\sqrt{\pi}} - \operatorname{diag}(D_{\sqrt{\pi}}\Psi^*\Psi D_{\sqrt{\pi}})\| \leq \|D_{\sqrt{\pi}}\Psi^*\Psi D_{\sqrt{\pi}}\| = \|\Psi D_{\sqrt{\pi}}\|^2.
$$

Further note that the k-th entry of $e_k^* H$ is zero, meaning
$$\|e_k^* H D_{\sqrt{\pi}}\| = \|(e_k^* - \psi_k^* \Psi) D_{\sqrt{\pi}}\| \leq \|\psi_k^* \Psi D_{\sqrt{\pi}}\| \leq \|\Psi D_{\sqrt{\pi}}\|.$$

So, as long as $\|\Psi D_{\sqrt{\pi}}\| \leq 1/3$, which holds in both regimes, we get
$$\|\mathbb{E}[D_{\sqrt{\pi}}^{-1} R_I^* H_{I,I} H_{I,I}^* R_I D_{\sqrt{\pi}}^{-1}]\| \leq 9/2 \cdot \|\Psi D_{\sqrt{\pi}}\|^4 + 3/2 \cdot \|\Psi D_{\sqrt{\pi}}\|^2$$
$$\leq (9/2 \cdot 1/3^2 + 3/2) \cdot \|\Psi D_{\sqrt{\pi}}\|^2 = 2\|\Psi D_{\sqrt{\pi}}\|^2.$$

In case $\mathit{H} = (\Psi \mathcal{E} - Z)^* \Psi = (\Phi \mathcal{E} - Z D_\alpha)^* \Psi$, we use
$$\|D_{\sqrt{\pi}} \mathit{H} D_{\sqrt{\pi}}\| \leq \|(\Phi \mathcal{E} - Z D_\alpha) D_{\sqrt{\pi}}\| \cdot \|\Psi D_{\sqrt{\pi}}\| = (\|\Phi D_{\sqrt{\pi}} \mathcal{E} - Z D_{\sqrt{\pi}} D_\alpha\|) \cdot \|\Psi D_{\sqrt{\pi}}\|$$
$$\leq (\|\Phi D_{\sqrt{\pi}}\| \cdot \varepsilon^2/2 + \|Z D_{\sqrt{\pi}}\|) \cdot \|\Psi D_{\sqrt{\pi}}\|.$$

Further since $\mathcal{E}_{kk} = \langle \psi_k, z_k \rangle$ and $\mathbb{I} - \psi_k \psi_k^*$ is an orthogonal projection we have
$$\|e_k^* \mathit{H} D_{\sqrt{\pi}}\| \leq \|\psi_k \langle \psi_k, z_k \rangle - z_k\| \cdot \|\Psi D_{\sqrt{\pi}}\|$$
$$= \|(\psi_k \psi_k^* - \mathbb{I}) z_k\| \cdot \|\Psi D_{\sqrt{\pi}}\| \leq \|z_k\| \cdot \|\Psi D_{\sqrt{\pi}}\| \leq \varepsilon \cdot \|\Psi D_{\sqrt{\pi}}\|.$$

So, using that $\varepsilon \leq \sqrt{2}$, as long as $\|\Phi D_{\sqrt{\pi}}\| \leq 1$, we get
$$2 \cdot \|\mathbb{E}[D_{\sqrt{\pi}}^{-1} R_I^* \mathit{H}_{I,I} \mathit{H}_{I,I}^* R_I D_{\sqrt{\pi}}^{-1}]\|$$
$$\leq (9\|\Phi D_{\sqrt{\pi}}\|^2 \cdot \varepsilon^4/4 + 9\varepsilon^2 \|Z D_{\sqrt{\pi}}\| + 9\|Z D_{\sqrt{\pi}}\|^2 + 3\varepsilon^2) \cdot \|\Psi D_{\sqrt{\pi}}\|^2.$$
$$\leq (9 \cdot \varepsilon^2/2 + 18\varepsilon\|Z D_{\sqrt{\pi}}\| + 9\|Z D_{\sqrt{\pi}}\|^2 + 3\varepsilon^2) \cdot \|\Psi D_{\sqrt{\pi}}\|^2$$
$$\leq (3\|Z D_{\sqrt{\pi}}\| + 3\varepsilon)^2 \cdot \|\Psi D_{\sqrt{\pi}}\|^2.$$

Finally note that $\|D_{\sqrt{\pi}} \mathit{H} D_{\sqrt{\pi}}\| = \|D_{\sqrt{\pi}} \mathit{H}^* D_{\sqrt{\pi}}\|$. Combining this with the bound
$$\|e_k^* \mathit{H}^* D_{\sqrt{\pi}}\| \leq \|\psi_k\| \cdot \|(\Phi \mathcal{E} - Z D_\alpha) D_{\sqrt{\pi}}\| \leq \|\Phi D_{\sqrt{\pi}}\| \cdot \varepsilon^2/2 + \|Z D_{\sqrt{\pi}}\|. \qquad (84)$$

we get that again as long as $\|\Psi D_{\sqrt{\pi}}\| \leq 1/3$
$$\|\mathbb{E}[D_{\sqrt{\pi}}^{-1} R_I^* \mathit{H}_{I,I}^* \mathit{H}_{I,I} R_I D_{\sqrt{\pi}}^{-1}]\| \leq (\|\Phi D_{\sqrt{\pi}}\| \cdot \varepsilon^2/2 + \|Z D_{\sqrt{\pi}}\|)^2 \cdot (9/2 \cdot \|\Psi D_{\sqrt{\pi}}\|^2 + 3/2)$$
$$\leq (\|\Phi D_{\sqrt{\pi}}\| \cdot \varepsilon^2/2 + \|Z D_{\sqrt{\pi}}\|)^2 \cdot 2,$$

which proves the third inequality. **(55-57)**✓

(c) We want to show that for a subset $\mathcal{G}$ of all supports of size S and a pair of $d \times K$ matrices $W = (w_1 \ldots, w_K)$ and $V = (v_1, \ldots, v_K)$ we have
$$\|\mathbb{E}[W R_I^* R_I V^* \cdot \mathbb{1}_I(\ell) \mathbb{1}_{\mathcal{G}}(I)]\| \leq \pi_\ell \cdot (\|W D_{\sqrt{\pi}}\| \cdot \|V D_{\sqrt{\pi}}\| + \|w_\ell\| \cdot \|v_\ell\|).$$

For $\ell \in I$ we can rewrite $W R_I^* R_I V^* = W \operatorname{diag}(\mathbf{1}_I) V^* = W \operatorname{diag}(\mathbf{1}_{I \setminus \{\ell\}}) V^* + w_\ell v_\ell^*$. Using this split and that by Theorem 15(c) we have $\mathbb{P}(\{\ell, k\} \subseteq I) \leq \pi_\ell \pi_k$ yields
$$\|\mathbb{E}[W R_I^* R_I V^* \cdot \mathbb{1}_I(\ell) \mathbb{1}_{\mathcal{G}}(I)]\|$$
$$\leq \|W \mathbb{E}\left[\operatorname{diag}(\mathbf{1}_{I \setminus \{\ell\}}) \cdot \mathbb{1}_I(\ell) \mathbb{1}_{\mathcal{G}}(I)\right] V^*\| + \|w_\ell v_\ell^*\| \cdot \mathbb{E}\left[\mathbb{1}_I(\ell) \mathbb{1}_{\mathcal{G}}(I)\right]$$
$$\leq \|W D_{\sqrt{\pi}}\| \cdot \|D_{\sqrt{\pi}}^{-1} \mathbb{E}\left[\operatorname{diag}(\mathbf{1}_{I \setminus \{\ell\}}) \cdot \mathbb{1}_I(\ell) \mathbb{1}_{\mathcal{G}}(I)\right] D_{\sqrt{\pi}}^{-1}\| \cdot \|D_{\sqrt{\pi}} V^*\| + \|w_\ell v_\ell^*\| \cdot \pi_\ell$$
$$= \|W D_{\sqrt{\pi}}\| \cdot \|V D_{\sqrt{\pi}}\| \cdot \max_{k \neq \ell} \pi_k^{-1} \cdot \mathbb{P}(\{\ell, k\} \subseteq I \cap I \in \mathcal{G}) + \|w_\ell\| \cdot \|v_\ell\| \cdot \pi_\ell$$
$$\leq \pi_\ell \cdot (\|W D_{\sqrt{\pi}}\| \cdot \|V D_{\sqrt{\pi}}\| + \|w_\ell\| \cdot \|v_\ell\|)$$

31

which completes the proof of (c). **(c)**✓

(d) We prove that for a general matrix H with zero diagonal we have

$$\|\mathbb{E}[D_{\sqrt{\pi}}^{-1}\mathbb{I}_{\ell^c}R_I^*H_{I,I}H_{I,I}^*R_I\mathbb{I}_{\ell^c}D_{\sqrt{\pi}}^{-1}\cdot\mathbb{1}_I(\ell)]\|$$
$$\leq \frac{\pi_\ell}{1-\pi_\ell}\left(3\|D_{\sqrt{\pi}}He_\ell\|^2 + \max_k H_{k\ell}^2 + \tfrac{9}{2}\|D_{\sqrt{\pi}}HD_{\sqrt{\pi}}\|^2 + \tfrac{3}{2}\max_k\|e_k^*HD_{\sqrt{\pi}}\|^2\right). \quad (85)$$

Using (83) and the identity $\mathbb{I}_{\ell^c} = \mathrm{diag}(\mathbf{1}_{[K]\setminus\{\ell\}})$, we first rewrite

$$\mathbb{I}_{\ell^c}R_I^*H_{I,I}H_{I,I}^*R_I\mathbb{I}_{\ell^c}\cdot\mathbb{1}_I(\ell) = (R_I^*H_{I,I}H_{I,I}^*R_I)\odot(\mathbf{1}_{[K]\setminus\{\ell\}}\mathbf{1}_{[K]\setminus\{\ell\}}^*)\cdot\mathbb{1}_I(\ell)$$
$$= \sum_k(H_kH_k^*)\odot\left(\mathbf{1}_{I\setminus\{k,\ell\}}\mathbf{1}_{I\setminus\{k,\ell\}}^*\cdot\mathbb{1}_I(\ell)\mathbb{1}_I(k)\right).$$

As before an application of the Schur Product Theorem and Theorem 15(d) leads to

$$\mathbb{E}_S[\mathbb{I}_{\ell^c}R_I^*H_{I,I}H_{I,I}^*R_I\mathbb{I}_{\ell^c}\cdot\mathbb{1}_I(\ell)] = \sum_k(H_kH_k^*)\odot\mathbb{E}_S\left[\mathbf{1}_{I\setminus\{k,\ell\}}\mathbf{1}_{I\setminus\{k,\ell\}}^*\cdot\mathbb{1}_I(\ell)\mathbb{1}_I(k)\right]$$

$$\preceq \frac{\pi_S(\ell)}{1-\pi_S(\ell)}(H_\ell H_\ell^*)\odot\mathbb{E}_{S-1}[\mathbf{1}_I\mathbf{1}_I^*]$$

$$\qquad\qquad + \sum_{k\neq\ell}\frac{\pi_S(\ell)}{1-\pi_S(\ell)}\left(H_k\frac{\pi_S(k)}{1-\pi_S(k)}H_k^*\right)\odot\mathbb{E}_{S-2}[\mathbf{1}_I\mathbf{1}_I^*]$$

$$\preceq \frac{\pi_S(\ell)}{1-\pi_S(\ell)}\left((H_\ell H_\ell^*)\odot\mathbb{E}_{S-1}[\mathbf{1}_I\mathbf{1}_I^*] + M\odot\mathbb{E}_{S-2}[\mathbf{1}_I\mathbf{1}_I^*]\right),$$

where again $M = H\,\mathrm{diag}(\frac{\pi_S}{1-\pi_S})H^*$. Finally, applying again Theorem 15 (e) and (b) and similar simplifications as in the proof of (b) yield

$$\|\mathbb{E}_S[D_{\sqrt{\pi}}^{-1}\mathbb{I}_{\ell^c}R_I^*H_{I,I}H_{I,I}^*R_I\mathbb{I}_{\ell^c}D_{\sqrt{\pi}}^{-1}\cdot\mathbb{1}_I(\ell)]\|$$
$$\leq \frac{\pi_\ell}{1-\pi_\ell}\left(\|(D_{\sqrt{\pi}}^{-1}H_\ell H_\ell^*D_{\sqrt{\pi}}^{-1})\odot\mathbb{E}_{S-1}[\mathbf{1}_I\mathbf{1}_I^*]\| + \|(D_{\sqrt{\pi}}^{-1}MD_{\sqrt{\pi}}^{-1})\odot\mathbb{E}_{S-2}[\mathbf{1}_I\mathbf{1}_I^*]\|\right)$$
$$\leq \frac{\pi_\ell}{1-\pi_\ell}\left(3\|D_{\sqrt{\pi}}H_\ell H_\ell^*D_{\sqrt{\pi}}\| + \|\mathrm{diag}(H_\ell H_\ell^*)\| + 3\|D_{\sqrt{\pi}}MD_{\sqrt{\pi}}\| + \|\mathrm{diag}(M)\|\right)$$
$$\leq \frac{\pi_\ell}{1-\pi_\ell}\left(3\|D_{\sqrt{\pi}}He_\ell\|^2 + \max_k H_{k\ell}^2 + \tfrac{9}{2}\|D_{\sqrt{\pi}}HD_{\sqrt{\pi}}\|^2 + \tfrac{3}{2}\max_k\|e_k^*HD_{\sqrt{\pi}}\|^2\right),$$

which together with our assumption that $\|\pi\|_\infty \leq 1/3$ leads to (d). **(d)** ✓

(73-75) Finally, we specialise the general inequality from (d) to the three concrete cases needed in the proofs of Lemma 7-10. Reusing the bounds for H from the corresponding special case of (b) as well as $H_{k\ell}^2 = |\langle\psi_k,\psi_\ell\rangle|^2 \leq \mu(\Psi)^2$ and the assumption $\|\Psi D_{\sqrt{\pi}}\| \leq 1/3$, yields

$$\|\mathbb{E}[D_{\sqrt{\pi}}^{-1}\mathbb{I}_{\ell^c}R_I^*H_{I,I}H_{I,I}^*R_I\mathbb{I}_{\ell^c}D_{\sqrt{\pi}}^{-1}\cdot\mathbb{1}_I(\ell)]\|$$
$$\leq 3/2\cdot\pi_\ell\cdot\left(3\cdot\|\Psi D_{\sqrt{\pi}}\|^2 + \mu(\Psi)^2 + 9/2\cdot1/9\cdot\|\Psi D_{\sqrt{\pi}}\|^2 + 3/2\cdot\|\Psi D_{\sqrt{\pi}}\|^2\right)$$
$$\leq 3/2\cdot\pi_\ell\cdot\max\{\mu(\Psi)^2,\|\Psi D_{\sqrt{\pi}}\|^2\}\cdot6 \leq 9\cdot\pi_\ell\cdot\max\{\mu(\Psi)^2,\|\Psi D_{\sqrt{\pi}}\|^2\}.$$

32

To prove the other two inequalities note that for $\mathit{И} = W^* V$ we get

$$\|\mathbb{E}[D_{\sqrt{\pi}}^{-1}\mathbb{I}_{\ell^c}R_I^*\mathit{И}_{I,I}\mathit{И}_{I,I}^*R_I\mathbb{I}_{\ell^c}D_{\sqrt{\pi}}^{-1}\cdot\mathbb{1}_I(\ell)]\|\cdot\frac{1-\pi_\ell}{\pi_\ell}$$

$$\leq 3\|D_{\sqrt{\pi}}W^*v_\ell\|^2 + \max_k(w_k^*v_\ell)^2 + 9\|D_{\sqrt{\pi}}W^*VD_{\sqrt{\pi}}\|^2 + 3\max_k\|w_k^*VD_{\sqrt{\pi}}\|^2$$

$$\leq \left(\max_k\|w_k\|^2 + 3\|WD_{\sqrt{\pi}}\|^2\right)\cdot\left(\max_\ell\|v_\ell\|^2 + 3\|VD_{\sqrt{\pi}}\|^2\right). \tag{86}$$

Applying this to $\mathit{И} = (\Psi\mathcal{E} - Z)^*\Psi$ and reusing the bounds for the corresponding special case of (b) with $\|\Psi D_{\sqrt{\pi}}\| \leq 1/8$ yields

$$\|\mathbb{E}[D_{\sqrt{\pi}}^{-1}\mathbb{I}_{\ell^c}R_I^*\mathit{И}_{I,I}\mathit{И}_{I,I}^*R_I\mathbb{I}_{\ell^c}D_{\sqrt{\pi}}^{-1}\cdot\mathbb{1}_I(\ell)]\|$$

$$\leq 3/2\cdot\pi_\ell\cdot\left[3\cdot(\|\Psi D_{\sqrt{\pi}}\|\cdot\varepsilon^2/2 + \|ZD_{\sqrt{\pi}}\|)^2 + \varepsilon^2\right]\cdot\left(1 + 3\cdot\|\Psi D_{\sqrt{\pi}}\|^2\right)$$

$$\leq 3/2\cdot\pi_\ell\cdot\delta^2\cdot\left[3\cdot(\|\Psi D_{\sqrt{\pi}}\|\cdot\varepsilon/2 + 1)^2 + 1\right]\cdot\left(1 + 3\cdot\|\Psi D_{\sqrt{\pi}}\|^2\right)\leq 9\cdot\pi_\ell\cdot\delta^2.$$

Due to the symmetry of (86) this also proves the case of $\mathit{И}^*$. **(73-75)**✓

∎

References

[1] A. Agarwal, A. Anandkumar, P. Jain, and P. Netrapalli. Learning sparsely used over-complete dictionaries via alternating minimization. *SIAM Journal on Optimization*, 26(4):2775–2799, 2016.

[2] A. Agarwal, A. Anandkumar, and P. Netrapalli. A clustering approach to learning sparsely used overcomplete dictionaries. *IEEE Transactions on Information Theory*, 63(1):575–592, 2017.

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

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

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

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

[7] 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.

[8] N. Chatterji and P. Bartlett. Alternating minimization for dictionary learning with random initialization. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, 2017.

[9] W. Dai and Y. Ye. A characterization on singular value inequalities of matrices. *Journal of Function Spaces*, 2020:1–4, 02 2020. doi: 10.1155/2020/1657381.

[10] D.L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4): 1289–1306, 2006.

[11] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407 – 499, 2004.

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

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

[19] 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.

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

[21] S. Minsker. On some extensions of Bernstein's inequality for self-adjoint operators. *Statistics and Probability Letters*, 127:111–119, 2017.

[22] B. Natarajan. Sparse approximate solutions to linear systems. *SIAM J. Computing*, 25(2):227–234, 1995.

[23] M.C. Pali and K. Schnass. Dictionary learning – from local towards global and adaptive. *Information and Inference: A Journal of the IMA*, 12(3), 2023.

[24] M.C. Pali, S. Ruetz, and K. Schnass. Average performance of OMP and Thresholding under dictionary mismatch. *IEEE Signal Processing Letters*, 29:1077–1081, 2022.

[25] 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.

[26] 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.

[27] 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.

[28] S. Ruetz. Adapted variable density subsampling for compressed sensing. *arXiv:2206.13796*, 2022.

[29] S. Ruetz. *Compressed Sensing and Dictionary Learning with Non-Uniform Support Distribution*. PhD thesis, University of Innsbruck, 2022.

[30] S. Ruetz. Convergence of ITKrM for dictionary learning. *in preparation*, 2023.

[31] 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.

[32] S. Ruetz and K. Schnass. Non-asymptotic bounds for inclusion probabilities in rejective sampling. *arXiv:2212.09391*, 2022.

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

[34] 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.

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

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

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

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

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

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

[41] A. M. Tillmann. On the computational intractability of exact and approximate dictionary learning. *IEEE Signal Processing Letters*, 22(1):45–49, 2015.

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