Signal Processing and Speech Communication Laboratory
Graz University of Technology
Inffeldgasse 12
AT-8010 Graz

Master Thesis

Sebastian Tschiatschek

Compressed Sensing of Block-Sparse Signals

Supervisors: MSc Patrick Kuppinger
MSc Graeme Pope
Dipl.-Ing. Dr. Christian Feldbauer

Assessor: Dipl.-Ing. Dr. Christian Feldbauer

March 2010

This thesis was conducted at:

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Communication-Theory Group
Prof. Dr. H. Bölcskei
Sternwartstrasse 7
CH-8092 Zürich

Abstract

Compressed Sensing is an alternative to Shannon-Nyquist sampling that provides techniques for sampling and reconstructing sparse signals at a rate far below the Nyquist rate. We consider compressed sensing under the model-assumption of block-sparsity, i.e., under the assumption that the nonzero coefficients of sparse signals appear in clusters.

A sufficient condition for having unique block-sparse representations in highly structured dictionaries, i.e., dictionaries that consist of the unions of orthonormal bases, is derived. This condition can allow for higher block-sparsities than conditions ignoring the dictionary structure.

For the same type of dictionaries we obtain a sufficient condition that guarantees the successful recovery, i.e., the perfect reconstruction, of any block k-sparse signal with a mixed ℓ_2/ℓ_1 optimization problem and with Block Orthogonal Matching Pursuit (BOMP). This condition can guarantee the exact recovery of signals with a higher block-sparsity than conditions derived for arbitrary dictionaries.

We also justify the usage of BOMP as an approximation algorithm to the block-sparse approximation problem in dictionaries with a small block Babel function. From this we conclude that the usage of BOMP can be advantageous over the usage of Orthogonal Matching Pursuit (OMP) if the signals to approximate exhibit block-structure. As a demonstration we show that BOMP can approximate signals that are block-sparse more accurate than OMP if they are observed through noisy measurements.

Zusammenfassung

Compressed Sensing ist eine Alternative zum Shannon-Nyquist-Sampling, die es erlaubt Signale, die in einer gegeben Basis bzw. einem gegeben Dictionary ("übervollständige Basis") eine spärliche Darstellung besitzen, weit unter ihrer Nyquist-Frequenz abzutasten und dennoch exakt zu rekonstruieren. In dieser Arbeit wird die Klasse von Signalen mit einer spärlichen bzw. näherungsweise spärlichen Darstellung betrachtet, in der die von Null verschiedenen Koeffizienten in Clustern bzw. Blöcken auftreten (Signale mit einer blockweise spärlichen Darstellung).

In Dictionaries ist die Darstellung eines Signals wegen der Übervollständigkeit im Allgemeinen nicht eindeutig. Für hochstrukturierte Dictionaries, die aus der Vereinigung mehrerer orthonormaler Basen bestehen, wird eine hinreichende Bedingung für eine eindeutige blockweise spärliche Signaldarstellung ermittelt. Diese Bedingung kann die eindeutige Darstellung von einer größeren Anzahl an Signalen gewährleisten als ähnliche Bedingungen für beliebige Dictionaries.

Für dieselben hochstrukturierten Dictionaries wird eine hinreichende Bedingung für die perfekte Rekonstruktion der blockweise spärlichen Darstellung eines Signals mittels Block Orthogonal Matching Pursuit (BOMP) und einem ℓ_2/ℓ_1-Norm Optimierungsproblem bestimmt. Diese Bedingung kann die Rekonstruktion der Darstellung von einer größeren Anzahl an Signalen gewährleisten als ähnliche Bedingungen für beliebige Dictionaries.

Betrachtet man schließlich beliebige Signale in schwach strukturierten Dictionaries (die einzelnen Blöcke der Dictionaries sind orthonormal), so können diese meist nicht exakt durch eine blockweise spärliche Darstellung (mit einer festgelegten maximalen Anzahl an von Null verschiedenen Blöcken) repräsentiert werden. Es ist jedoch oft möglich, diese Signale durch eine blockweise spärliche Darstellung gut zu approximieren. Es wird gezeigt, dass eine derartige Approximation mittels BOMP ermittelt werden kann, wenn das betrachtete Dictionary eine langsam wachsende Block-Babel-Funktion aufweist. Hieraus folgt, dass die Verwendung von BOMP vorteilhaft gegenüber der Verwendung von Orthogonal Matching Pursuit (OMP) sein kann, wenn die Darstellung des zu approximierenden Signals eine Block-Struktur aufweist. Zur Demonstration wird gezeigt, dass die Darstellung derartiger Signale aus verrauschten Messungen mittels BOMP exakter rekonstruiert werden kann als mittels OMP.

Statutory Declaration

I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

March 18, 2010

Signature

Acknowledgements

I would like to thank Patrick Kuppinger and Graeme Pope for supervising my thesis at ETH and introducing me to the very interesting field of Compressed Sensing. Furthermore, I would like to thank Christian Feldbauer for supervising my thesis at TU-Graz. Moreover, I am very grateful to the Austrian scholarship service as well as ETH for providing the funds for my stay in Zurich.

Contents

Chapter 1

Introduction

1.1 Introduction

To perform discrete time signal processing we often need to sample data from the "real world". Following the classical Shannon-Nyquist sampling theorem, one has to uniformly sample a signal at a rate of at least twice its bandwidth to be able to recover it exactly [1], [2]. In many cases, the obtained signal is compressed right after acquisition for efficient storage. This two step process is quite resource intensive and wasteful in terms of necessary sensing resources, processing power and interim storage needs. Think for example of digital cameras that acquire a large set of pixels, typically in the millions, and then compress the image drastically for storage reasons using some standard like JPEG2000 [3], [4].

For the class of sparse signals, we can combine the sampling and the compression process by applying the paradigm of *compressed sensing*. It suggests to take a few linear measurements (samples)—far less samples than in the case of Shannon-Nyquist sampling—of the signal of interest and recover it later by appropriate nonlinear methods. This works because the linear measurements combine the information of the underlying signal into the individual samples [5]. In contrast, when applying Shannon-Nyquist sampling on a sparse signal one measures lots of zero samples. Hence, compressed sensing combines the processes of sampling and compression, avoiding oversampling [6].

The paradigm of compressed sensing is applicable widely, since many natural signals have an almost sparse (compressible) representation in appropriate dictionaries and can be sufficiently approximated by sparse signals [7]. These dictionaries may be predetermined, e.g., the union of the spike- and discrete Fourier-basis, or machine learned as in [8] for image compression.

The nonzero coefficients of signals of some classes exhibit further structure and do not occur at arbitrary positions but according to some rule. Such a rule could be, that the nonzero coefficients occur in clusters (blocks). This specific rule is covered by the *block-sparsity*-model [9], [10].

The block-sparsity model is not just a theoretical model as block-sparse vectors appear naturally in multiple measurement vectors [11] or when dealing with multiband signals [12]. Because of the relevance of this model we investigated if, and how, some results on unique sparse representation and signal recovery for highly structured dictionaries carry over from the regular sparse case to the block-sparse case [13], [14]. By highly structured dictionaries we mean dictionaries that consist of orthonormal blocks or the unions of orthonormal bases.

We do not carry out a stochastic analysis as for example in [15]. Instead, we make a deterministic investigation of the fundamental problem in compressed sensing, that is, the recovery of an unknown vector from an under-determined system of linear equations and derive our results from there.

1.2 Contributions and Main Results

This thesis contributes improved conditions for unique sparse representation and exact signal recovery for highly structured dictionaries in compressed sensing under the block-sparsity assumption. Furthermore, it justifies the usage of Block Orthogonal Matching Pursuit (BOMP, [16]) as an approximation algorithm to the block-sparse approximation problem. In particular the contributions of this thesis include:

- A sufficient condition on the block-sparsity for unique block-sparse representations in dictionaries that are unions of orthonormal bases. This condition can be advantageous to similar conditions obtained without taking the special structure of the dictionary into account.

- A sufficient condition on the block-sparsity of signals in dictionaries comprised of orthonormal bases that ensures that they can be recovered by mixed ℓ_2/ℓ_1-norm optimization (L-OPT) and BOMP. Again, an advantage over similar conditions for arbitrary dictionaries is possible.

- A bound on the maximal approximation error of an approximation for an arbitrary signal obtained by BOMP in k-steps in relation to the approximation error of an optimal k-block approximation. From this the usage of BOMP as an approximation algorithm to the block-sparse approximation problem in dictionaries with a small block Babel function is justified.

1.3 Thesis Overview

In Chapter 2 we give an overview of compressed sensing for the regular sparse case, including conditions for unique sparse representations and signal recovery methods. The same material is presented for compressed sensing in the block-sparse case.

Chapter 3 covers the derivation of the mentioned improved conditions that mainly carry over from the regular sparse case. For all derivations we provide a theoretical discussion that emphasizes the advantage over the sparse case and demonstrate it by simulations.

The justification for the usage of BOMP as an approximation algorithm to the block-sparse approximation problem is covered in Chapter 4. Additionally, a bound on the maximal approximation error is derived. We provide a theoretical discussion and demonstrate the usage of BOMP as an approximation algorithm by simulations.

Lemmas, Corollaries and Propositions stated without proof are taken from the papers or books mentioned at the beginning of the statements. All other statements are due to our own work. Statements that have the note “adopted from ...” at their beginning, closely follow the mentioned reference but needed some manipulation or enhanced arguments to carry over to the block-sparse case.

The discussions and simulations on our contributions are our own work, otherwise this is mentioned explicitly.

1.4 Notation

Throughout this thesis we use small bold face letters like $\mathbf{x}$ to denote vectors and capital bold face letters like $\mathbf{A}$ to denote matrices. The ith element of a vector $\mathbf{x}$ is x_i and the (i,j)th element of a matrix $\mathbf{A}$ is denoted as $A_{i,j}$.

A comprehensive list of symbols and operators used throughout this thesis can be found in the Appendix, Section A.

1.4.1 Matrices

The matrix $\mathbf{I}_N$ is the identity matrix of size $N \times N$. When there is no ambiguity we often denote it simply as $\mathbf{I}$.

The symbol $\mathbf{0}_N$ represents either the all zero matrix of size $N \times N$ or the all zero vector of length N, depending on the context. If it is clear which value N takes, we simply write $\mathbf{0}$.

The kernel $\ker(\mathbf{A})$ of a matrix $\mathbf{A} \in \mathbb{C}^{M\times N}$ is the set

$$\ker(\mathbf{A}) := \left\{\mathbf{x} \in \mathbb{C}^N : \mathbf{A}\mathbf{x} = \mathbf{0}\right\}. \tag{1.1}$$

1.4.2 Block Indexing

Despite in Chapter 2.1 we assume the dimensions of vectors and matrices to be integer multiples of some block-size d (which is also an integer). Then we index parts of the considered vectors and matrices as described below.

Consider some vector $\mathbf{x} \in \mathbb{C}^N$, with $N := Rd$, where the number of blocks R in the vector is an integer. This vector is partitioned in blocks as

$$\mathbf{x} = [\underbrace{x_1 \ \dots \ x_d}_{\mathbf{x}^T[1]} \ \underbrace{x_{d+1} \ \dots \ x_{2d}}_{\mathbf{x}^T[2]} \ \dots \ \underbrace{x_{N-d+1} \ \dots \ x_N}_{\mathbf{x}^T[R]}]^T. \tag{1.2}$$

That is, we denote the ith block of $\mathbf{x}$ as $\mathbf{x}[i]$.

Similarly, we partition some matrix $\mathbf{A} \in \mathbb{C}^{M\times N}$, with $M := Qd$ and $N := Rd$, where both Q and R are integers and d is the block-size, as

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}[1,1] & \mathbf{A}[1,2] & \cdots & \mathbf{A}[1,R] \\ \mathbf{A}[2,1] & \mathbf{A}[2,2] & \cdots & \mathbf{A}[2,R] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{A}[Q,1] & \mathbf{A}[Q,2] & \cdots & \mathbf{A}[Q,R] \end{bmatrix}, \tag{1.3}$$

where

$$\mathbf{A}[l,r] = \begin{bmatrix} A_{l',r'} & A_{l',r'+1} & \dots & A_{l',r'+d-1} \\ A_{l'+1,r'} & A_{l'+1,r'+1} & \dots & A_{l'+1,r'+d-1} \\ \vdots & \vdots & \ddots & \vdots \\ A_{l'+d-1,r'} & A_{l'+d-1,r'+1} & \dots & A_{l'+d-1,r'+d-1} \end{bmatrix}, \tag{1.4}$$

and $l' = (l-1)d + 1$ and $r' = (r-1)d + 1$.

In this partition $\mathbf{A}[l, r] \in \mathbb{C}^{d \times d}$. We call $\mathbf{A}[l, r]$ the (l, r)th $d \times d$ submatrix of $\mathbf{A}$, or simply the (l, r)th submatrix of $\mathbf{A}$ if it is clear which block-size we refer to.

We also index a matrix $\mathbf{A} \in \mathbb{C}^{P \times N}$, where N is as before, only by one index, i.e., $\mathbf{A}[l]$. Then, the matrix $\mathbf{A}$ is partitioned as

$$\mathbf{A} = \left[\begin{array}{cccc} \mathbf{A}[1] & \mathbf{A}[2] & \cdots & \mathbf{A}[R] \end{array} \right], \tag{1.5}$$

with $\mathbf{A}[l] \in \mathbb{C}^{P \times d}$. We call $\mathbf{A}[l]$ the lth block of matrix $\mathbf{A}$

1.4.3 Norms and Mixed Norms

We state the definition of the ℓ_p-norm of vectors and matrices [17] as a reference for the mixed norms:

Definition 1.1 (Vector Norms). *For $p \geq 1$ the ℓ_p-norm of the vector $\mathbf{x} \in \mathbb{C}^N$ is*

$$\|\mathbf{x}\|_p := \left(\sum_{i=1}^{N} |x_i|^p \right)^{1/p}. \tag{1.6}$$

Definition 1.2 (Induced Matrix Norm). *The induced matrix norm $\|\mathbf{A}\|_p$ is*

$$\|\mathbf{A}\|_p := \max_{\mathbf{x}, \mathbf{x} \neq \mathbf{0}} \frac{\|\mathbf{A}\mathbf{x}\|_p}{\|\mathbf{x}\|_p}, \tag{1.7}$$

with $\|\mathbf{z}\|_p$ being the ℓ_p-norm of the vector $\mathbf{z}$.

By the above definition the induced ℓ_1 matrix norm of a matrix $\mathbf{A} \in \mathbb{C}^{M \times N}$ becomes

$$\|\mathbf{A}\|_1 = \max_{1 \leq j \leq N} \sum_{i=1}^{M} |A_{i,j}|, \tag{1.8}$$

i.e., it is the maximum absolute column sum of $\mathbf{A}$ (see [17]). The induced ℓ_∞ matrix norm of $\mathbf{A}$ can be calculated as

$$\|\mathbf{A}\|_\infty = \max_{1 \leq i \leq M} \sum_{j=1}^{N} |A_{i,j}|, \tag{1.9}$$

i.e., it equals the maximum absolute row sum of $\mathbf{A}$ (see [17]).

Another matrix norm that we will use regularly is the spectral norm. For a matrix $\mathbf{A}$ the spectral norm is denoted as $\rho(\mathbf{A})$ and can be calculated as

$$\rho(\mathbf{A}) = \sqrt{\sigma_{\max}(\mathbf{A}^H \mathbf{A})}, \tag{1.10}$$

where $\sigma_{\max}(\mathbf{B})$ denotes the largest eigenvalue of the positive-semidefinite matrix $\mathbf{B}$ (see [17]).

We can carry the concept of norms over to account for the block-structure of vectors and matrices used in most of our derivations by introducing the mixed vector norms and the mixed matrix norms. For this we assume the integer d to be the block-size.

Definition 1.3 (Mixed Vector Norm [16]). *Let $\mathbf{x} \in \mathbb{C}^N$ be a vector with $N = Rd$, where R is an integer and d is the block-size. Define the mixed vector norm $\|\mathbf{x}\|_{2,p}$ of $\mathbf{x}$ to be*

$$\|\mathbf{x}\|_{2,p} := \|\mathbf{v}\|_p, \tag{1.11}$$

where $v_i := \|\mathbf{x}[i]\|_2$ for $i = 1, \ldots, R$.

Definition 1.4 (Mixed Matrix Norm [16]). *The mixed matrix norm $\|\mathbf{A}\|_{2,p}$ of a matrix $\mathbf{A} \in \mathbb{C}^{M \times N}$, with $M = Qd$ and $N = Rd$, where Q and R are both integers and d is the block-size, is*

$$\|\mathbf{A}\|_{2,p} := \max_{\mathbf{x}, \mathbf{x} \neq \mathbf{0}} \frac{\|\mathbf{A}\mathbf{x}\|_{2,p}}{\|\mathbf{x}\|_{2,p}}. \tag{1.12}$$

To obtain similar results for mixed matrix norms as in (1.8) and (1.9), that will appear in some derivations later, we introduce the quantities $\rho_c(\cdot)$ and $\rho_r(\cdot)$ as follows:

Definition 1.5 (From [16], Lemma 1). *Let $\mathbf{A} \in \mathbb{C}^{M,N}$ be some matrix with $M = Qd$ and $N = Rd$, where Q, R, d are integers and d is the block-size. Define $\rho_c(\mathbf{A})$ and $\rho_r(\mathbf{A})$ as*

$$\rho_r(\mathbf{A}) := \max_l \sum_r \rho(\mathbf{A}[l,r]), \text{ and} \tag{1.13}$$

$$\rho_c(\mathbf{A}) := \max_r \sum_l \rho(\mathbf{A}[l,r]). \tag{1.14}$$

With this we can relate the mixed matrix norms $\|\mathbf{A}\|_{2,1}$ and $\|\mathbf{A}\|_{2,\infty}$ to $\rho_c(\mathbf{A})$ and $\rho_r(\mathbf{A})$ as covered by the following lemma:

Lemma 1.1 (From [16], Lemma 1). *Let $\mathbf{A} \in \mathbb{C}^{M \times N}$ be a matrix, with $M = Qd$ and $N = Rd$, where d is the block-size and d, Q, R are integers. Then,*

$$\|\mathbf{A}\|_{2,\infty} \leq \rho_r(\mathbf{A}), \text{ and} \tag{1.15}$$

$$\|\mathbf{A}\|_{2,1} \leq \rho_c(\mathbf{A}). \tag{1.16}$$

For our derivations it is important to note that $\rho_c(\cdot)$ as defined above is a matrix norm [16]. This is covered by the following lemma.

Lemma 1.2 (From [16], Lemma 2)**.** $\rho_c(\mathbf{A})$ *as introduced in Definition 1.5 is a matrix norm and as such satisfies the following properties: For any two matrices* $\mathbf{A}$ *and* $\mathbf{B}$ *(of appropriate dimensions),*

1. $\rho_c(\mathbf{A}) = 0$ *if and only if* $\mathbf{A} = \mathbf{0}$ *(positive definiteness),*
2. $\rho_c(\alpha\mathbf{A}) = |\alpha|\rho_c(\mathbf{A})$ *for all* $\alpha \in \mathbb{C}$ *(homogeneity),*
3. $\rho_c(\mathbf{A} + \mathbf{B}) \leq \rho_c(\mathbf{A}) + \rho_c(\mathbf{B})$ *(triangle inequality), and*
4. $\rho_c(\mathbf{AB}) \leq \rho_c(\mathbf{A})\, \rho_c(\mathbf{B})$ *(submultiplicity).*

Chapter 2

Compressed Sensing

In this chapter we describe some fundamentals of compressed sensing. Further we give an overview of some refinements resulting from incorporating the block-sparsity model.

2.1 Fundamentals

Compressed sensing describes a mean of data acquisition for sparse signals. In contrast to classical Shannon-Nyquist sampling one does not uniformly sample the signal of interest at an appropriate rate, but takes only a few linear measurements. Recovery can then be carried out by exploiting the signal structure, i.e., its sparsity.

From this, the underlying problem in compressed sensing is the recovery of an unknown vector $\mathbf{x}$, from a number of linear measurements $\mathbf{y}$ that is much smaller than the dimension of $\mathbf{x}$. Because of the linearity of the measurements, they can be described by a measurement matrix $\mathbf{D}$ according to $\mathbf{y} = \mathbf{D}\mathbf{x}$. Hence, compressed sensing resorts to solving an underdetermined system of linear equations. Later we will refer to $\mathbf{D}$ as a dictionary.

If $\mathbf{x}$ is sufficiently sparse in a dictionary $\mathbf{D}$, then this will allow us to recover $\mathbf{x}$ exactly from the measurements $\mathbf{y}$.

As a beginning we describe the sparse representation and sparse approximation problem that build the foundation of compressed sensing. Then we dive into dictionary characterization. This will give rise to properties like the coherence of a dictionary—these properties will appear naturally in various estimates, as for example conditions under which sparse representations are unique. Finally, we consider the problem of recovering the unknown vector $\mathbf{x}$ from the measurements. Again we will use dictionary properties to give a characterization of the recovery algorithms.

2.1.1 Sparse Representation

The main idea behind compressed sensing is that many natural signals have *sparse representations* in some basis, or, more generally, in some *dictionary* [18].

Definition 2.1 (Dictionary)**.** *A dictionary for the space* $\mathbb{C}^M$ *is a set of nonzero vectors* $\{\mathbf{d}_i\}$*, where* $i \in \Delta := \{1, 2, \ldots, N\}$*, that span the space* $\mathbb{C}^M$*. The vectors* $\mathbf{d}_i$ *are also called atoms.*

We usually consider the matrix version of the dictionary, i.e.,

$$\mathbf{D} = [\mathbf{d}_1\, \mathbf{d}_2\, \ldots\, \mathbf{d}_N] \in \mathbb{C}^{M \times N}. \tag{2.1}$$

Thus, a dictionary is, informally, the generalization of a basis of some space to a possibly redundant description of the same space. Since the columns of $\mathbf{D}$ are required to span $\mathbb{C}^M$, we have $N \geq M$.

We assume that all dictionaries have normalized atoms, i.e., $\|\mathbf{d}_i\|_2 = 1$ for all $i \in \Delta$.

A *representation* of a signal $\mathbf{y}$ in a dictionary $\mathbf{D}$ is given by a vector $\mathbf{x}$, so that $\mathbf{y} = \mathbf{D}\mathbf{x}$. For a vector to be *k-sparse* it means, that is has at most k nonzero coefficients.

Definition 2.2 (Sparsity)**.** *A vector* $\mathbf{x} \in \mathbb{C}^N$ *is k-sparse, if*

$$|\mathrm{supp}(\mathbf{x})| \leq k. \tag{2.2}$$

We say that $\mathbf{y}$ *is k-sparse in a dictionary* $\mathbf{D}$*, if there exists a vector* $\mathbf{x}$ *such that* $\mathbf{y} = \mathbf{D}\mathbf{x}$*, where* $\mathbf{x}$ *is k-sparse. For ease of notation we will write* $\|\mathbf{x}\|_0$ *for* $|\mathrm{supp}(\mathbf{x})|$*.*

Now we give an informal description of the *sparse representation problem.* We consider a signal

$$\mathbf{y} \in \{\mathbf{D}\hat{\mathbf{x}} : \|\hat{\mathbf{x}}\|_0 \leq k\}, \tag{2.3}$$

where $\mathbf{D}$ is a fixed dictionary and k a fixed positive integer. The problem is to find the sparsest representation $\mathbf{x}$ of $\mathbf{y}$ in the dictionary $\mathbf{D}$. In general there is no unique sparsest representation. However, if we limit the maximal sparsity k of $\hat{\mathbf{x}}$ in (2.3) sufficiently, then the sparsest representation of any possible $\mathbf{y}$ becomes unique—which we will clarify later.

Note that decreasing k in (2.3) can only decrease the number of possible signals $\mathbf{y}$.

Example 2.1. *Consider the dictionary*

$$\mathbf{D} = \begin{bmatrix} 1 & 0 & 0 & 1/\sqrt{3} \\ 0 & 1 & 0 & 1/\sqrt{3} \\ 0 & 0 & 1 & 1/\sqrt{3} \end{bmatrix}.$$

The signal

$$\mathbf{y}_1 = [1\ 1\ 0]^T \in \{\mathbf{D}\hat{\mathbf{x}} : \|\hat{\mathbf{x}}\|_0 \leq 2\}$$

has no unique sparsest representation, as for example

$$\begin{aligned}\mathbf{x}_1 &= [1\ 1\ 0\ 0]^T\ , \ \textit{and}\\ \mathbf{x}_1' &= \left[0\ 0\ -1\ \sqrt{3}\right]^T\end{aligned}$$

are sparsest representations.

In contrast, the signal

$$\mathbf{y}_2 = [1\ 0\ 0]^T \in \{\mathbf{D}\hat{\mathbf{x}} : \|\hat{\mathbf{x}}\|_0 \leq 1\}$$

has the unique sparsest representation

$$\mathbf{x}_2 = [1\ 0\ 0\ 0]^T\ .$$

Finally, we state the sparse representation problem formally.

Problem 2.1 (Sparse Representation)**.** *Given a dictionary* $\mathbf{D}$ *of size* $M \times N$*, some positive integer* k *and a vector*

$$\mathbf{y} \in \{\mathbf{D}\hat{\mathbf{x}} : \|\hat{\mathbf{x}}\|_0 \leq k\}, \tag{2.4}$$

find the representation $\mathbf{x} \in \mathbb{C}^N$ *as the solution to the minimization problem*

$$\arg\min_{\mathbf{x}\in\mathbb{C}^N} \|\mathbf{x}\|_0 \quad \textit{s.t. } \mathbf{y} = \mathbf{D}\mathbf{x}. \tag{2.5}$$

At the end of this section, we want to emphasize that a well chosen dictionary can allow for the sparse representation of certain classes of signals. This is demonstrated in the following example.

Example 2.2 (Sparse Representation in Dictionaries)**.** *Consider the discrete signal* $\mathbf{y} \in \mathbb{C}^M$ *given by*

$$y_i = \sin\left(\frac{2\pi}{M}(i-1)\right) + \delta[i-1], \quad i = 1, \ldots, M,$$

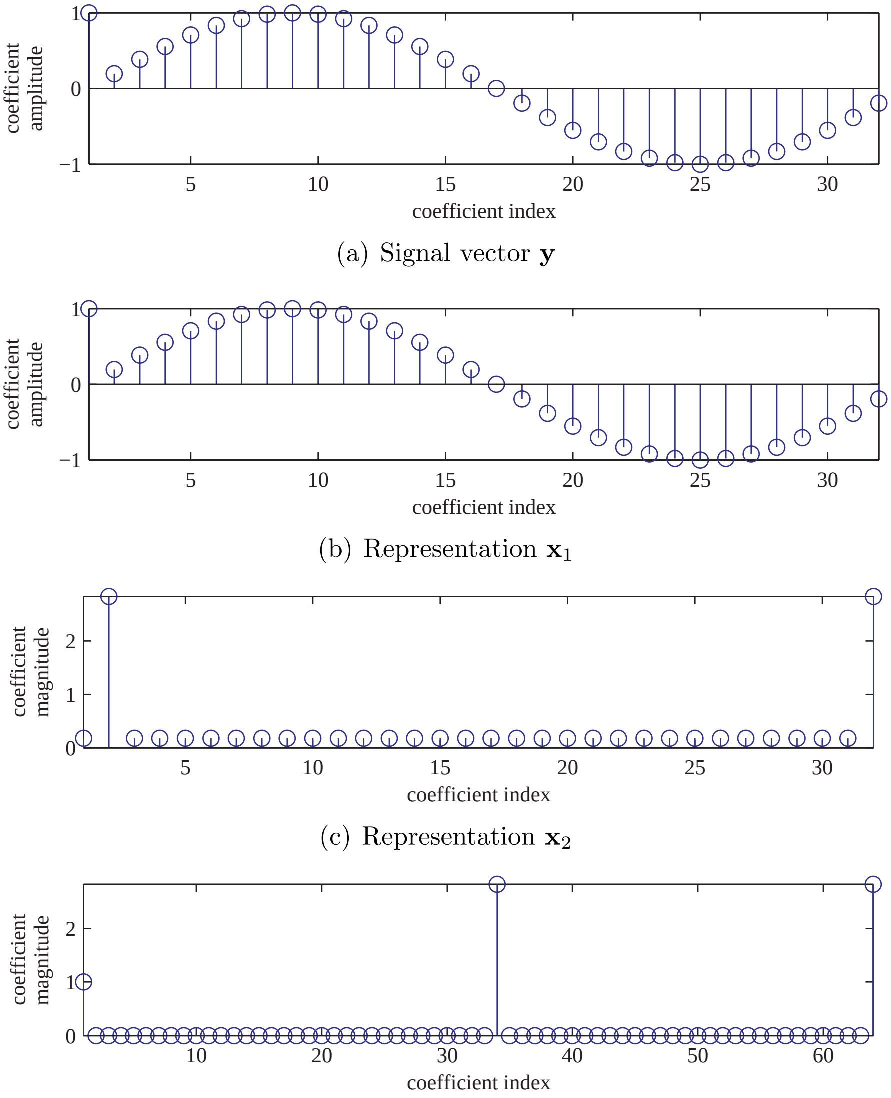
where $\delta[\cdot]$ *is the Kronecker delta and* $M = 32$*.*

If we want to represent $\mathbf{y}$ *in the dictionary* $\mathbf{D}_1 = \mathbf{I}_M$*, then the support of the representation* $\mathbf{x}_1 \in \mathbb{C}^M$ *such that* $\mathbf{y} = \mathbf{D}\mathbf{x}_1$ *has cardinality* $M-1$*.*

On the other hand, if we choose the dictionary to be $\mathbf{D}_2 = \mathbf{F}$*, where* $\mathbf{F}$ *denotes the DFT matrix of size* $M \times M$*, i.e.,* $F_{l,r} = (1/\sqrt{M})\exp(i2\pi lr/M)$*, then the representation* $\mathbf{x}_2 \in \mathbb{C}^M$ *such that* $\mathbf{y} = \mathbf{D}\mathbf{x}_2$ *has exactly* M *nonzero coefficients.*

Finally, if we choose $\mathbf{D}$ *to be* $\mathbf{D} = [\mathbf{I}\ \mathbf{F}]$*, then the sparsest representation* $\mathbf{x} \in \mathbb{C}^{2M}$ *such that* $\mathbf{y} = \mathbf{D}\mathbf{x}$ *has only three nonzero coefficients.*

This example is illustrated in Figure 2.1 for $M = 32$ *and* $j = 1$*.*



(a) Signal vector $\mathbf{y}$

(b) Representation $\mathbf{x}_1$

(c) Representation $\mathbf{x}_2$

(d) Representation $\mathbf{x}_3$

Figure 2.1: Illustration of Example 2.2. (a) Signal $\mathbf{y}$. (b) Representation $\mathbf{x}_1$ of $\mathbf{y}$ in the $\mathbf{D}_1 = \mathbf{I}$. (This plot equals the signal $\mathbf{y}$ itself.). (c) Representation $\mathbf{x}_2$ of $\mathbf{y}$ in $\mathbf{D}_2 = \mathbf{F}$ (the coefficient magnitudes are shown). (d) Sparsest representation $\mathbf{x}_3$ of $\mathbf{y}$ in $\mathbf{D} = [\mathbf{I}\ \mathbf{F}]$. (the coefficient magnitudes are shown).

2.1.2 Sparse Approximation

We again consider the problem of representing some signal $\mathbf{y}$ by a sparse vector $\mathbf{x}$ in a dictionary $\mathbf{D}$ of size $M \times N$. In contrast to before, we do not put any constraints on the signal $\mathbf{y}$. That is, $\mathbf{y}$ can be any element in the space $\mathbb{C}^M$.

Whenever the sparsity level of $\mathbf{x}$ is less than the dimensionality M, we cannot represent all vectors $\mathbf{y}$ exactly. We thus change our goal to find the *best k*-sparse approximation. The quality of an approximation $\mathbf{a}$ is measured by the ℓ_2-norm of the difference of $\mathbf{y}$ and the approximation itself, i.e., by $\|\mathbf{y} - \mathbf{a}\|_2$. That is, the quality measure is the energy of the approximation error (we simply refer to this as the approximation error). The lower this approximation error is, the better is the considered approximation. Thus, the sparse approximation problem is as follows.

Problem 2.2 (Sparse Approximation)**.** *Given a dictionary* $\mathbf{D}$ *of size* $M \times N$ *and a vector* $\mathbf{y} \in \mathbb{C}^M$ *find the k-sparse vector* $\mathbf{x}_{opt}$ *as the solution to the minimization problem*

$$\arg \min_{\mathbf{x} \in \mathbb{C}^N} \|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2 \quad s.t. \ \|\mathbf{x}\|_0 \leq k. \tag{2.6}$$

We call the approximation $\mathbf{a}_{opt} = \mathbf{D}\mathbf{x}_{opt}$ *an "optimal k-term approximation of* $\mathbf{y}$ *in* $\mathbf{D}$*".*

The solution to the previous problem is not necessarily unique. That is, there may be distinct k-sparse vectors $\mathbf{x}_{\text{opt}}$ that minimize the value of the objective function in (2.6). If this is the case, we will be interested in finding any one of these optimal approximations.

2.1.3 Dictionary Characterization

As mentioned before, we can make the solution to the sparse representation problem unique for all possible signals $\mathbf{y}$ by limiting the sparsity level k that is allowed for $\hat{\mathbf{x}}$ in Equation (2.3). We call a limit on k that guarantees uniqueness of the problem a *sufficient uniqueness condition.*

One generally wants that such a limit is as high as possible. The higher this limit the more signals have a unique sparsest representation. The best limit on k depends on the considered dictionary and is in general hard do determine. However, we can give estimates based on properties characterizing the dictionary. These properties are the *coherence* and the *Babel function*, which we will now introduce.

Definition 2.3 (Coherence)**.** *Let* $\mathbf{D}$ *be a dictionary consisting of* N *atoms* $\mathbf{d}_i$*, with* $i \in \Delta$*. The coherence* $\mu(\mathbf{D})$ *of the dictionary is defined as*

$$\mu(\mathbf{D}) := \max_{\substack{i,j \in \Delta \\ i \neq j}} |\langle \mathbf{d}_i, \mathbf{d}_j \rangle|. \tag{2.7}$$

In other words, the coherence measures how similar any two distinct atoms of the dictionary are. We write μ instead of $\mu(\mathbf{D})$ when it is clear, which dictionary is considered. By the above definition, the coherence of an orthogonal basis is 0. Also note that because of the normalized atoms of the dictionary the coherence obeys $0 \leq \mu \leq 1$.

A way to characterize a dictionary in more detail than by the coherence is the Babel function, which we define as follows.

Definition 2.4 (Babel Function [19]). *Let $\mathbf{D}$ be a dictionary as in Definition 2.3. The Babel function $\mu_1(\mathbf{D}, k)$ of the dictionary is defined as*

$$\mu_1(\mathbf{D}, k) := \max_{|\Delta'|=k} \max_{i} \sum_{j \in \Delta'} |\langle \mathbf{d}_i, \mathbf{d}_j \rangle|, \tag{2.8}$$

where $\Delta' \subset \Delta$ and $i \in (\Delta \backslash \Delta')$.

The Babel function measures the maximal sum of the absolute values of the inner products between a fixed atom and k distinct atoms of the dictionary. When there is no danger of confusion, we write $\mu_1(k)$ instead of $\mu_1(\mathbf{D}, k)$.

Note the following facts about the Babel function: $\mu_1(1) = \mu$, and $\mu_1(k) \leq k\,\mu_1(1) = k\,\mu$.

2.1.4 Unique Sparse Representation

We are now in the position to state conditions such that the sparse representation problem has a unique solution.

Proposition 2.1 (From [14], Theorem 1 and Corollary 1). *For arbitrary dictionaries $\mathbf{D}$ the solution to the sparse representation problem is unique, if*

$$\mathbf{y} \in \{\mathbf{D}\hat{\mathbf{x}} : \|\hat{\mathbf{x}}\|_0 \leq k\},$$

where k satisfies

$$k < \frac{1}{2}\left(1 + \frac{1}{\mu}\right). \tag{2.9}$$

As a special case, for dictionaries $\mathbf{D}$ that are the unions of L orthonormal bases, the sparse representation problem has a unique solution if

$$k < \frac{1}{2}\left(1 + \frac{1}{L-1}\right)\frac{1}{\mu}. \tag{2.10}$$

Furthermore, if $\mathbf{y} = \mathbf{D}\mathbf{x}$ and $\|\mathbf{x}\|_0 \leq k$, for some k satisfying (2.9) *(or* (2.10) *if $\mathbf{D}$ is the union of L orthonormal bases), then $\mathbf{x}$ is the unique solution to the sparse representation problem.*

Note that if the dictionary $\mathbf{D}$ in the above proposition consists of the union of L orthonormal bases, then both bounds (2.9) and (2.10) are applicable. In this case one could state the sufficient condition on k to guarantee uniqueness of the sparse representation problem as

$$k < \max\left\{\frac{1}{2}\left(1+\frac{1}{\mu}\right), \frac{1}{2}\left(1+\frac{1}{L-1}\right)\frac{1}{\mu}\right\}.$$

2.1.5 Signal Recovery Methods

Consider the problem of actually solving the sparse representation problem. That is, we want to recover the representation $\mathbf{x}$ from $\mathbf{y} = \mathbf{Dx}$, where $\mathbf{x}$ is as sparse as possible. We assume to have full knowledge of $\mathbf{D}$. The basic idea is to solve the optimization problem

$$\arg\min_{\mathbf{x}\in\mathbb{C}^N} \|\mathbf{x}\|_0 \quad \text{s.t. } \mathbf{y} = \mathbf{Dx}. \tag{2.11}$$

This problem will always recover a vector $\mathbf{x}$ that has minimal sparsity. Furthermore, if the recovered vector has sparsity satisfying Proposition 2.1, the recovered vector will be the unique sparsest representation of $\mathbf{y}$. However, solving the problem (2.11) resorts to a combinatorial search, which is in general an NP-hard problem and thus not feasible.

To overcome this difficulty, recovery algorithms, with different complexity and performance, have been invented. We only mention *Basis Pursuit* (BP) [20] and *Orthogonal Matching Pursuit* (OMP) [21] here, since they will appear later in several investigations.

Basis Pursuit

The underlying idea of BP is to find the solution to (2.11) by solving the minimization problem

$$\arg\min_{\mathbf{x}\in\mathbb{C}^N} \|\mathbf{x}\|_1 \quad \text{s.t. } \mathbf{y} = \mathbf{Dx}. \tag{2.12}$$

For all signals that have a k-sparse representation in $\mathbf{D}$ the solutions to (2.11) and (2.12) are guaranteed to be unique and to coincide, when k is sufficiently small—conditions for this are summarized in Section 2.1.6. The advantage of replacing problem (2.11) by (2.12), is that the minimization of the ℓ_1-norm can be solved by linear programming. Thus, the optimization problem can be solved efficiently using interior point methods or methods that make use of the underlying sparsity as in [22].

Orthogonal Matching Pursuit

OMP [21] is a greedy algorithm that tries to recover $\mathbf{x}$ from $\mathbf{y} = \mathbf{Dx}$. It works as follows: Let $\mathbf{r}_j$ be the residual and $\mathbf{a}_j$ the approximation of $\mathbf{y}$ obtained by OMP after the jth step. The set $\mathcal{I}$ is a set of indices of atoms of the dictionary and $\mathbf{x}_j$ is a j-sparse vector created by OMP in the jth step.

OMP initializes $\mathcal{I}$ as the empty set and sets $\mathbf{r}_0 = \mathbf{y}$.

In the jth step, OMP adds the index i_j to the index set $\mathcal{I}$, which is determined by solving

$$i_j = \arg\max_i |\langle \mathbf{d}_i, \mathbf{r}_{j-1} \rangle|. \tag{2.13}$$

In other words, i_j is the index of the atom that best approximates the residual. At this point $\mathcal{I}$ contains the indices of atoms chosen by OMP up to and including step j. Then, $\mathbf{x}_j$ is determined by solving the least squares problem

$$\mathbf{x}_j = \arg\min_{\mathbf{x}_j} \left\| \mathbf{y} - \sum_{i \in \mathcal{I}} \mathbf{d}_i x_{j,i} \right\|_2 , \tag{2.14}$$

where $x_{j,i}$ is the ith coefficient of the vector $\mathbf{x}_j$ and all $x_{j,l} = 0$ for $l \notin \mathcal{I}$.

Then the approximation $\mathbf{a}_j$ of $\mathbf{y}$ is calculated as

$$\mathbf{a}_j = \mathbf{D}\mathbf{x}_j. \tag{2.15}$$

Finally, the residual is set to

$$\mathbf{r}_j = \mathbf{y} - \mathbf{a}_j. \tag{2.16}$$

Some conditions that guarantee that OMP recovers the representation $\mathbf{x}$ from the signal $\mathbf{y} = \mathbf{D}\mathbf{x}$ exactly are summarized in Section 2.1.6.

2.1.6 Recovery Conditions

Basis Pursuit and Orthogonal Matching Pursuit recover any vector $\mathbf{x}$ from $\mathbf{y} = \mathbf{D}\mathbf{x}$ and $\mathbf{D}$ exactly in k steps, if $\mathbf{y}$ is the linear combination of at most k atoms of the dictionary and the *Exact Recovery Condition* is obeyed. This is captured by the following proposition.

Lemma 2.1 (Exact Recovery Condition; from [19], Theorem 3.1)**.** *Let* $\mathbf{x} \in \mathbb{C}^N$ *be a k-sparse vector,* $\mathbf{D}$ *a dictionary of size* $M \times N$ *and* $\mathbf{y} = \mathbf{D}\mathbf{x}$*. A sufficient condition for BP and OMP to recover* $\mathbf{x}$ *exactly in k steps is*

$$\|\mathbf{D}_{opt}^{\dagger} \overline{\mathbf{D}}_{opt}\|_1 < 1, \tag{2.17}$$

where $\mathbf{D}_{opt}$ *is a matrix consisting of the atoms of the dictionary that correspond to the nonzero coefficients of* $\mathbf{x}$ *and* $\mathbf{A}^{\dagger}$ *is the pseudoinverse of* $\mathbf{A}$*. The matrix* $\overline{\mathbf{D}}_{opt}$ *is comprised of all atoms in* $\mathbf{D}$*, that are not in* $\mathbf{D}_{opt}$*.*

In general one does not know in advance which atoms belong to the sparsest representation of some signal $\mathbf{y}$. However, there are sufficient conditions similar to the bounds in Proposition 2.1 that guarantee that the Exact Recovery Condition will hold.

Proposition 2.2 (From [19], Theorem B)**.** *The Exact Recovery Condition in Lemma 2.1 holds for every linear combination of k atoms from a dictionary $\mathbf{D}$ with coherence μ and Babel function $\mu_1(\cdot)$, whenever*

$$k < \frac{1}{2}\left(1 + \frac{1}{\mu}\right), \tag{2.18}$$

or, whenever

$$\mu_1(k-1) + \mu_1(k) < 1. \tag{2.19}$$

As a special case, if $\mathbf{D}$ is comprised of L orthonormal bases and

$$k < \left(\sqrt{2} - 1 + \frac{1}{2\,(L-1)}\right)\frac{1}{\mu}, \tag{2.20}$$

then the Exact Recovery Condition is obeyed.

2.1.7 Sparse Approximation with OMP

OMP is an interesting algorithm since it is easy to analyze and has provable approximation error bounds for signals that are not sparse, as the following proposition demonstrates.

Proposition 2.3 (From [19], Corollary 4.3)**.** *Let $\mathbf{y} \in \mathbb{C}^M$ be an arbitrary signal and $\mathbf{D}$ a dictionary of size $M \times N$ with Babel function $\mu_1(\cdot)$. If $\mu_1(k) < 1/2$, then OMP generates a k-term approximation $\mathbf{a}_k \in \mathbb{C}^M$ in k-steps, which satisfies*

$$\|\mathbf{y} - \mathbf{a}_k\|_2 \leq \sqrt{1 + \frac{k\,[1 - \mu_1(k)]}{[1 - 2\,\mu_1(k)]^2}}\,\|\mathbf{y} - \mathbf{a}_{opt}\|_2, \tag{2.21}$$

where $\mathbf{a}_{opt} \in \mathbb{C}^M$ is an optimal k-term approximation of $\mathbf{y}$.

Hence, whenever the condition from the above proposition is satisfied, we are guaranteed to obtain an approximation $\mathbf{a}_k$ by OMP in k-steps that lies in a hypersphere around the vector $\mathbf{y}$. The radius r of this hypersphere is at most the ℓ_2-distance between $\mathbf{y}$ and $\mathbf{a}_{opt}$ multiplied by the factor

$$c = \sqrt{1 + \frac{k\,[1 - \mu_1(k)]}{[1 - 2\,\mu_1(k)]^2}}. \tag{2.22}$$

This is illustrated for the two dimensional case in Figure 2.2.

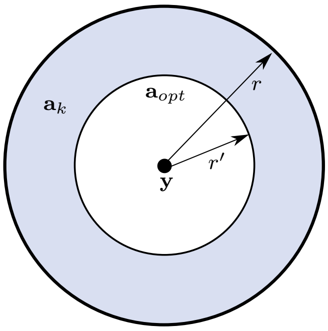


Figure 2.2: Illustration of Proposition 2.3 for the two dimensional case. The optimal k-sparse approximation $\mathbf{a}_{opt}$ of $\mathbf{y}$ in a dictionary $\mathbf{D}$ lies somewhere on the inner circle, while the obtained approximation $\mathbf{a}_k$ by OMP in k steps lies in the gray shaded region. The radius r equals $c\,r'$, with c as in (2.22) and $r' = \|\mathbf{y} - \mathbf{a}_{opt}\|_2$.

2.2 Block-Sparsity

Block-sparsity is a model assumption on the class of sparse signals. In a block-sparse signal, the nonzero coefficients do not occur at arbitrary positions, but in clusters. This assumption is of practical relevance, since block-sparse vectors arise naturally in multiple measurement vector (MMV) problems and when dealing with multi-band signals [11], [12].

We give a precise description of block-sparsity and describe how to carry the concepts of the earlier sections over to block-sparse signals. We put our focus on unique signal representation and signal recovery methods.

2.2.1 Block-Sparse Representation

As in the sparse case, we consider a vector $\mathbf{y}$ and its representation $\mathbf{x}$ in a dictionary $\mathbf{D}$ of size $M \times N$ such that $\mathbf{y} = \mathbf{D}\mathbf{x}$. Here and in the rest of this thesis we assume that N is an integer multiple of some fixed block-size d, that is $N = Rd$, where R and d are integers.

In general there can be multiple representations of a signal $\mathbf{y} \in \mathbb{C}^M$. We want to state conditions under which there is a unique *block-sparsest* representation—details to come later.

Block-Sparsity

Consider the vector $\mathbf{x} \in \mathbb{C}^N$. We say, that the ith block of $\mathbf{x}$ is nonzero, if $\|\mathbf{x}[i]\|_2 > 0$. The block-sparsity of a vector is the number of its nonzero blocks. To state this formally we introduce the *indicator function* $I(\cdot)$.

Definition 2.5 (Indicator Function, [16]). *The indicator function $I : \mathbb{R}_+ \to \{0, 1\}$ is*

$$I(a) := \begin{cases} 0, & \text{if } a = 0, \\ 1, & \text{otherwise.} \end{cases} \tag{2.23}$$

Definition 2.6 (Block-Sparsity, [16]). *The block-sparsity of a vector $\mathbf{x} \in \mathbb{C}^N$, where $N = Rd$ and d is the block-size, denoted $\|\mathbf{x}\|_{2,0}$, is*

$$\|\mathbf{x}\|_{2,0} := \sum_{l=1}^{R} I(\|\mathbf{x}[l]\|_2). \tag{2.24}$$

We say that a vector $\mathbf{x}$ is block k-sparse if $\|\mathbf{x}\|_{2,0} \le k$.

Now we give the definitions of two sets that will be important in our considerations and derivations.

Definition 2.7. *For a given length $N = Rd$, where R and d are integers and d is the block-size, define*

$$\mathcal{X}_k := \left\{\mathbf{x} \in \mathbb{C}^N : \|\mathbf{x}\|_{2,0} \le k\right\}. \tag{2.25}$$

In other words, the set $\mathcal{X}_k$ is the set of all block k-sparse vectors of length N.

Definition 2.8. *For a given dictionary $\mathbf{D}$ of size $M \times N$, define*

$$\mathcal{Y}_k := \mathbf{D}\mathcal{X}_k = \{\mathbf{D}\mathbf{x} : \mathbf{x} \in \mathcal{X}_k\}. \tag{2.26}$$

That is, $\mathcal{Y}_k$ is the set of all signals that have a block k-sparse representation in $\mathbf{D}$.

Similar to before, we define the block-sparse representation problem.

Problem 2.3 (Block-Sparse Representation). *Given a dictionary $\mathbf{D}$ of size $M \times N$, a positive integer k and a vector $\mathbf{y} \in \mathcal{Y}_k$, find the representation $\mathbf{x}$ as the solution to the minimization problem*

$$\arg\min_{\mathbf{x} \in \mathbb{C}^N} \|\mathbf{x}\|_{2,0} \quad s.t. \ \mathbf{y} = \mathbf{D}\mathbf{x}. \tag{2.27}$$

For small enough k the block-sparse representation problem will yield a unique solution for all possible signals $\mathbf{y}$—details to come later.

2.2.2 Block-Sparse Approximation

Assume that we want to represent some arbitrary vector $\mathbf{y} \in \mathbb{C}^M$ by a block k-sparse signal $\mathbf{x}$ in a dictionary $\mathbf{D}$. As this is in general not possible, we want to find the *optimal* k-block approximation of $\mathbf{y}$. An approximation $\mathbf{a}_{\text{opt}} = \mathbf{D}\mathbf{x}_{\text{opt}}$ is optimal if $\|\mathbf{y} - \mathbf{D}\mathbf{x}_{\text{opt}}\|_2$ achieves the minimum possible value of $\|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2$, where $\mathbf{x}$ is any block k-sparse vector. We define this to be the block-sparse approximation problem:

Problem 2.4 (Block-Sparse Approximation)**.** *Given a dictionary $\mathbf{D}$ of size $M \times N$ and an arbitrary vector $\mathbf{y} \in \mathbb{C}^M$ find the block k-sparse representation $\mathbf{x}_{opt}$ as the solution to the minimization problem*

$$\arg \min_{\mathbf{x} \in \mathbb{C}^N} \|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2 \quad s.t. \; \|\mathbf{x}\|_{2,0} \leq k. \tag{2.28}$$

We call the approximation $\mathbf{a}_{opt} = \mathbf{D}\mathbf{x}_{opt}$ "optimal k-block approximation of $\mathbf{y}$ in $\mathbf{D}$".

2.2.3 Dictionary Characterization

As in the regular sparse case we want to find quantities with the same relevance, as, for example, the coherence, that allow us to characterize the dictionary $\mathbf{D}$.

First we extend the way in which we index parts of a dictionary $\mathbf{D}$. In the earlier sections we used the index set Δ and its elements $i \in \Delta$ to address individual atoms $\mathbf{d}_i$ of the dictionary $\mathbf{D}$. We now additionally make use of the index set $\Delta_B := \{1, \dots, R\}$ to identify blocks of the dictionary.

We can now extend the concept of the coherence to the block case. That is, we introduce the *block-coherence* of a dictionary [16].

Definition 2.9 (Block-Coherence [16])**.** *Let $\mathbf{D}$ be a dictionary consisting of blocks $\mathbf{D}[l]$, with $l \in \Delta_B$. The block-coherence $\mu_B(\mathbf{D})$ of the dictionary is defined as*

$$\mu_B(\mathbf{D}) := \frac{1}{d} \max_{\substack{i,j \in \Delta_B \\ i \neq j}} \rho\Big(\mathbf{D}[i]^H \, \mathbf{D}[j]\Big). \tag{2.29}$$

Note that the block-coherence μ_B reduces to the coherence μ of a dictionary for the case $d = 1$.

We define the block Babel function in the same spirit as in the sparse case.

Definition 2.10 (Block Babel Function)**.** *Let $\mathbf{D}$ be a dictionary as in Definition 2.9. The block Babel function $\mu_{B1}(\mathbf{D}, k)$ of the dictionary is defined as*

$$\mu_{B1}(\mathbf{D}, k) := \frac{1}{d} \max_{|\Delta'_B| = k} \max_{i} \sum_{j \in \Delta'_B} \rho\Big(\mathbf{D}[i]^H \, \mathbf{D}[j]\Big), \tag{2.30}$$

with $\Delta'_B \subset \Delta_B$ and $i \in (\Delta_B \backslash \Delta'_B)$.

Note the following facts about the block Babel function: $\mu_{B1}(1) = \mu_B$, and $\mu_{B1}(k) \leq k\,\mu_{B1}(1) = k\,\mu_B$.

For ease of reading we say that a dictionary $\mathbf{D}$ consists of orthonormal blocks if the blocks of the dictionary itself are orthonormal, i.e., if for all $l \in \Delta_B$, we have

$$\mathbf{D}[l]^H\,\mathbf{D}[l] = \mathbf{I}_d.$$

In parts of this thesis we will consider dictionaries that are the union of L orthonormal bases, where each basis is of size $M \times M$. For these dictionaries we assume the block-size d to divide the dimension of the bases M.

2.2.4 Unique Block-Sparse Representation

We continue our discussion on the uniqueness of the representation $\mathbf{x}$ of $\mathbf{y}$ in a dictionary $\mathbf{D}$. As mentioned, there are in general many representations of $\mathbf{y}$ in $\mathbf{D}$. However, we can guarantee that there is unique block-sparsest representation by requiring $\mathbf{y} \in \mathcal{Y}_k$, where k is sufficiently small.

The following two propositions state sufficient conditions on k to guarantee this. Before stating them, there is one rather important fact about dictionaries to mention. The definition of a dictionary allows for linearly dependent vectors to occur within a single block. If this is the case, we cannot guarantee the unique block-sparse representation of any signal $\mathbf{y}$. To overcome this, we assume from now on that all dictionaries that are considered in a block sense, consist of blocks with linearly independent atoms.

Proposition 2.4 (Adopted from [14], Theorem 1, for a proof refer to the Appendix, Section C.3)**.** *Let $\mathbf{D}$ be a dictionary consisting of orthonormal blocks of size d and with block-coherence μ_B. If $\mathbf{y} \in \mathcal{Y}_k$ and*

$$k < \frac{1}{2}\left(1 + \frac{1}{d\,\mu_B}\right), \tag{2.31}$$

then the block-sparse representation problem has a unique solution.

Furthermore, $\mathbf{x}$ is the unique solution to the block-sparse representation problem if $\mathbf{y} = \mathbf{D}\mathbf{x}$ and $\|\mathbf{x}\|_{2,0} \leq k$.

Proposition 2.5 (Adopted from [19], for a proof see Appendix C.2)**.** *Let $\mathbf{D}$ be a dictionary comprised of orthonormal blocks of size d and with block Babel function $\mu_{B1}(m)$. If $\mathbf{y} = \mathbf{D}\mathbf{x}$, where $\mathbf{x}$ is block k-sparse, and*

$$k < \frac{1}{2}\min\{m : d\,\mu_{B1}(m-1) \geq 1\}, \tag{2.32}$$

then $\mathbf{x}$ is the block-sparsest representation of $\mathbf{y}$.

Consequently, for every $\mathbf{y} \in \mathcal{Y}_k$ with $k < 1/2\min\{m : d\,\mu_{B1}(m-1) \geq 1\}$, the block-sparse representation problem has a unique solution.

2.2.5 Signal Recovery Methods

To solve the block-sparse representation problem for some vector $\mathbf{y}$ in $\mathbf{D}$ one could try to evaluate the optimization problem

$$\arg\min_{\mathbf{x}\in\mathbb{C}^N} \|\mathbf{x}\|_{2,0} \quad \text{s.t. } \mathbf{y} = \mathbf{D}\mathbf{x}. \tag{2.33}$$

However, solving this problem requires a combinatorial search and is thus infeasible.

To tackle this problem, we can use recovery algorithms such as L-OPT [11], which is a mixed ℓ_2/ℓ_1-norm minimization, or Block Orthogonal Matching Pursuit (BOMP) [16]. These algorithms reduce the complexity of the problem and guarantee recovery of the underlying representation under certain conditions—details to come later. We will briefly describe these algorithms.

L-OPT [10]

L-OPT is the equivalent to Basis Pursuit for the block-sparse case. The idea behind this algorithm is to replace the optimization problem (2.33) by the problem

$$\arg\min_{\mathbf{x}\in\mathbb{C}^N} \|\mathbf{x}\|_{2,1} \quad \text{s.t. } \mathbf{y} = \mathbf{D}\mathbf{x}. \tag{2.34}$$

This new optimization problem can be cast as a second order cone program (SOCP) and hence can be solved efficiently [23], [10]. To make use of this optimization problem, we need to know under which conditions the solutions of (2.33) and (2.34) are unique and coincide—see Section 2.2.6 for details.

A variation of L-OPT is L-OPT-O. This algorithm orthonormalizes the blocks in the dictionary before solving (2.34). That is, it generates the dictionary $\mathbf{D}'$ from $\mathbf{D}$ such that

$$\begin{aligned} &\operatorname{span}\big(\mathbf{D}'[l]\big) = \operatorname{span}(\mathbf{D}[l]) \text{ and} \\ &\mathbf{D}'[l]^H\, \mathbf{D}'[l] = \mathbf{I}_d \end{aligned}$$

for all $l \in \Delta_B$ and then solves

$$\arg\min_{\mathbf{x}\in\mathbb{C}^N} \|\mathbf{x}'\|_{2,1} \quad \text{s.t. } \mathbf{y} = \mathbf{D}'\mathbf{x}'.$$

From the solution $\mathbf{x}'$ it finally calculates $\mathbf{x}$ by transforming the coordinates of $\mathbf{x}'[l]$ in the coordinate system given by the columns of $\mathbf{D}'[l]$ to the corresponding coordinates in the the coordinate system given by the columns of $\mathbf{D}[l]$ for all $l \in \Delta_B$. This transformation is unique if and only if the atoms within each block of the dictionary $\mathbf{D}$ are linearly independent—which holds for our dictionaries by assumption.

Block Orthogonal Matching Pursuit [16]

Block Orthogonal Matching Pursuit (BOMP) is the extension of OMP to the block-sparse case and works as follows: Let $\mathbf{r}_j$ be the residual and $\mathbf{a}_j$ an approximation obtained by BOMP after the jth step. Further, let $\mathcal{I}$ be a set of indices of blocks of the dictionary selected by the algorithm. Initially set $\mathbf{r}_0 = \mathbf{y}$ and $\mathcal{I}$ to the empty set.

In step j, BOMP selects the block from the dictionary best matched to the residual $\mathbf{r}_j$, i.e., the block with index

$$i_j = \arg\max_i \|\mathbf{D}[i]^H \, \mathbf{r}_{j-1}\|_2. \tag{2.35}$$

This index is then added to the set $\mathcal{I}$. Hence, at this point $\mathcal{I}$ contains the block-indices chosen by BOMP up to and including step j.

Then, the vector $\mathbf{x}_j$ is obtained as the solution to the least squares problem

$$\arg\min_{\mathbf{x}_j} \left\| \mathbf{y} - \sum_{i \in \mathcal{I}} \mathbf{D}[i] \, \mathbf{x}_j[i] \right\|_2, \tag{2.36}$$

where $\mathbf{x}_j[l] = \mathbf{0}$ for all $l \notin \mathcal{I}$. From this, the approximation $\mathbf{a}_j = \mathbf{D}\mathbf{x}_j$ and the residual $\mathbf{r}_j = \mathbf{y} - \mathbf{a}_j$ are determined.

There are conditions that guarantee that BOMP will successfully recover the representation $\mathbf{x}$ of $\mathbf{y} = \mathbf{D}\mathbf{x}$. See Section 2.2.6 for details.

A variation of BOMP is BOMP-O (it is a similar variation as L-OPT-O of L-OPT, compare this to Section 2.2.5).

2.2.6 Recovery Conditions

Lemma 2.2 (From [16], Theorem 2, Exact Recovery Condition)**.** *Let* $\mathbf{D}$ *be a dictionary and* $\mathbf{y} = \mathbf{D}\mathbf{x}$*, where* $\mathbf{x}$ *is block* k*-sparse. A sufficient condition for L-OPT and BOMP to recover* $\mathbf{x}$ *exactly is that*

$$\rho_c\left(\mathbf{D}_{opt}^{\dagger} \overline{\mathbf{D}}_{opt}\right) < 1, \tag{2.37}$$

where $\mathbf{D}_{opt}$ *is a matrix whose blocks correspond to the nonzero blocks of* $\mathbf{x}$ *in* $\mathbf{D}$ *and* $\overline{\mathbf{D}}_{opt}$ *consists of the blocks of* $\mathbf{D}$ *that are not in* $\mathbf{D}_{opt}$*.*

The condition in the above lemma is hard to evaluate since one does not know in advance which blocks of $\mathbf{x}$ are nonzero. However, we can find conditions on the block-sparsity k of $\mathbf{x}$ that ensure that Equation (2.37) holds, as in Proposition 2.6.

Proposition 2.6 (Adopted from [16], Theorem 1, for the proof see Appendix C.3)**.** *Let* $\mathbf{D}$ *be a dictionary consisting of orthonormal blocks and* $\mathbf{y} = \mathbf{D}\mathbf{x}$*, where* $\mathbf{x}$ *is block* k*-sparse. If*

$$k < \frac{1}{2}\left(1 + \frac{1}{d\,\mu_B}\right), \tag{2.38}$$

then the Exact Recovery Condition is obeyed. Hence, $\mathbf{x}$ *can be recovered by L-OPT and BOMP.*

Chapter 3

Refined Conditions

3.1 Sparsity in Unions of Orthonormal Bases

Consider the block-sparse representation problem in the dictionary $\mathbf{D}$, i.e., the problem of finding the block-sparsest representation of some signal $\mathbf{y} \in \mathcal{Y}_k$. It is then important to know for which values of k the solution to the problem is unique for all possible $\mathbf{y}$.

An answer to this question for sparse signals and dictionaries that are unions of orthonormal bases was derived in [14] and is summarized in Section 2.1.4. We generalize their results to block-sparse vectors by incorporating the block-sparsity assumption. In this way we obtain a sufficient condition on k that guarantees the uniqueness of the block-sparse representation problem for a potentially higher sparsity. This condition is our main result and summarized by the following proposition:

Proposition 3.1 (Adopted from [14], Corollary 1)**.** *Let* $\mathbf{D}$ *be a dictionary consisting of* L *orthonormal bases with block-size* d *and let* $\mathbf{y} \in \mathcal{Y}_k$*. If*

$$k < \frac{1}{2}\left(1 + \frac{1}{L-1}\right)\frac{1}{d\,\mu_B}, \tag{3.1}$$

then the block-sparse representation problem has a unique solution.

Furthermore, $\mathbf{x}$ *is the unique solution to the block-sparse representation problem, if* $\mathbf{y} = \mathbf{D}\mathbf{x}$ *and* $\|\mathbf{x}\|_{2,0} \leq k$*.*

3.1.1 Derivation

Note that the derivation closely follows the path taken in [14].

We start by defining some quantities for vectors and dictionaries in a block sense.

Definition 3.1 (Block-Support)**.** *The block-support of a vector* $\mathbf{x} \in \mathbb{C}^N$ *with* $N = Rd$*, where* R *and* d *are integers and* d *is the block-size, is*

$$supp_B(\mathbf{x}) := \{l : \|\mathbf{x}[l]\|_2 > 0\}. \tag{3.2}$$

In other words, the block-support of a vector is the set of indices of the nonzero blocks of $\mathbf{x}$.

Definition 3.2 (Block-Spark)**.** *The block-spark of a dictionary* $\mathbf{D}$ *is*

$$Z_B(\mathbf{D}) := \min_{\substack{\mathbf{z}\in\ker(\mathbf{D})\\ \mathbf{z}\neq\mathbf{0}}} \|\mathbf{z}\|_{2,0}. \tag{3.3}$$

That is, the block-spark is the smallest number of blocks of the dictionary such that the vectors within these blocks are linearly dependent. Note that finding the block-spark of a dictionary is a combinatorial problem and thus in general not feasible. At this point we remind the reader of the assumption made in Section 2.2.4 that the atoms within each block of the dictionary are linearly independent.

We can now state a sufficient condition for the minimization problem

$$\arg\min_{\mathbf{x}\in\mathbb{C}^N} \sum_l I(\|\mathbf{x}[l]\|_2) \quad \text{s.t. } \mathbf{y} = \mathbf{D}\mathbf{x}, \tag{3.4}$$

to have a unique solution.

Note that the objective function of the above minimization problem equals the block-sparsity of $\mathbf{x}$. Hence, the solution to the problem minimizes the block-sparsity of the vector $\mathbf{x}$, under the constraint $\mathbf{y} = \mathbf{D}\mathbf{x}$, and in this way solves the block-sparse representation problem.

Lemma 3.1 (Adopted from [14], Lemma 1)**.** *Let* $\mathbf{D}$ *be a dictionary consisting of* R *blocks and* $S \subset \Delta_B = \{1, \ldots, R\}$ *be a set of block-indices. Define*

$$P_{B,0}(S, \mathbf{D}) := \max_{\substack{\mathbf{z}\in\ker(\mathbf{D})\\ \mathbf{z}\neq\mathbf{0}}} \frac{\sum\limits_{l\in S} I(\|\mathbf{z}[l]\|_2)}{\sum\limits_{l} I(\|\mathbf{z}[l]\|_2)}. \tag{3.5}$$

If $P_{B,0}(S, \mathbf{D}) < 1/2$*, then, for all* $\mathbf{x}$ *such that* $\mathbf{y} = \mathbf{D}\mathbf{x}$ *and* $supp_B(\mathbf{x}) \subset S$*,* $\mathbf{x}$ *is the unique solution to the problem* (3.4)*.*

Proof. To show that the solution of the minimization problem (3.4) is unique under the given assumptions with minimizer $\mathbf{x}$, we have to prove that any vector $\mathbf{x}' \neq \mathbf{x}$ with $\mathbf{y} = \mathbf{D}\mathbf{x}'$ results in a higher value of the objective function in (3.4) and is thus not the solution to the problem. Note that we can express $\mathbf{x}'$ as $\mathbf{x}' = \mathbf{x} + \mathbf{z}'$, where $\mathbf{z}'$ is in the kernel of $\mathbf{D}$.

Let $\mathbf{x}$ be such that $\mathbf{y} = \mathbf{D}\mathbf{x}$ and $\text{supp}_B(\mathbf{x}) \subset S$. Further, assume that $P_{B,0}(S, \mathbf{D}) < 1/2$. Then, for any $\mathbf{z} \in \ker(\mathbf{D})$, we have that

$$\sum_{l\notin S} I(\|\mathbf{z}[l]\|_2) - \sum_{l\in S} I(\|\mathbf{z}[l]\|_2) > 0.$$

Observe that the indicator function $I(\cdot)$ and the composed function $I(\|\cdot\|_2)$ obey the triangle inequality. Thus, $I(\|\cdot\|_2)$ also obeys the reverse triangle inequality, i.e., $I(\|\mathbf{a}+\mathbf{b}\|_2) - I(\|\mathbf{a}\|_2) \geq -I(\|\mathbf{b}\|_2)$. Hence, we get

$$\sum_{l \notin S} I(\|\mathbf{z}[l]\|_2) + \sum_{l \in S} \left[I(\|\mathbf{x}[l] + \mathbf{z}[l]\|_2) - I(\|\mathbf{x}[l]\|_2)\right] > 0.$$

The support of $\mathbf{x}$ is a subset of S by assumption. Hence, we can rewrite the inequality as

$$\sum_{l} I(\|\mathbf{x}[l] + \mathbf{z}[l]\|_2) > \sum_{l} I(\|\mathbf{x}[l]\|_2).$$

The above inequality holds for any nonzero $\mathbf{z} \in \ker(\mathbf{D})$. Therefore, $\mathbf{x}$ is the unique minimizer of the problem (3.4). □

The next lemma gives a sufficient condition on the cardinality of the set S in Lemma 3.4 to ensure that $P_{B,0}(S, \mathbf{D})$ is less than 1/2.

Lemma 3.2 (Adopted from [14], Lemma 2)**.** *Let S and $\mathbf{D}$ be as before. If*

$$|S| < \lceil Z_B(\mathbf{D})\,/2 \rceil\,, \tag{3.6}$$

then $P_{B,0}(S, \mathbf{D}) < 1/2$.

Proof. Assume that $|S| < \lceil Z_B(\mathbf{D})\,/2\rceil$. We show that

$$\frac{\sum\limits_{l \in S} I(\|\mathbf{z}[l]\|_2)}{\sum\limits_{l} I(\|\mathbf{z}[l]\|_2)} < 1/2 \tag{3.7}$$

for any nonzero $\mathbf{z} \in \ker(\mathbf{D})$.

To see this, note that the numerator of (3.7) is bounded above by $|S|$, and is thus at most $\lceil Z_B(\mathbf{D})\,/2\rceil - 1$ by assumption. On the other hand, the minimum value the denominator can take is $Z_B(\mathbf{D})$ by the definition of the block-spark. Thus we have

$$\begin{aligned}\frac{\sum\limits_{l \in S} I(\|\mathbf{z}[l]\|_2)}{\sum\limits_{l} I(\|\mathbf{z}[l]\|_2)} &\leq \frac{\lceil Z_B(\mathbf{D})\,/2\rceil - 1}{Z_B(\mathbf{D})} \\ &< \frac{Z_B(\mathbf{D})\,/2}{Z_B(\mathbf{D})} \\ &= \frac{1}{2}.\end{aligned}$$

Hence, the statement follows. □

As a next step we give an estimate of the block-spark for dictionaries that are comprised of orthonormal bases. Combined with Lemma 3.2 this allows us to state a sufficient condition on the block-sparsity k in the block-sparse representation problem that guarantees uniqueness of its solution.

Lemma 3.3 (Adopted from [14], Lemma 3)**.** *Let* $\mathbf{D}$ *be a dictionary with block-coherence* μ_B *comprised of* L *orthonormal bases such that*

$$\mathbf{D} = [\mathbf{B}_1\, \mathbf{B}_2\, \dots\, \mathbf{B}_L],$$

where $\mathbf{B}_i \in \mathbb{C}^{M\times M}$ *are orthonormal bases,* $M = Q'd$, Q' *and* d *are integers, and* d *is the block-size.*

Then, the block-spark of the dictionary $\mathbf{D}$ *satisfies*

$$Z_B(\mathbf{D}) \geq \left(1 + \frac{1}{L-1}\right) \frac{1}{d\,\mu_B}. \tag{3.8}$$

Proof. Assume that $\mathbf{z} \in \ker(\mathbf{D})$. The vector $\mathbf{z}$ can be partitioned as

$$\mathbf{z} = \begin{bmatrix} \mathbf{z}_1 \\ \vdots \\ \mathbf{z}_L \end{bmatrix},$$

with $\mathbf{z}_l \in \mathbb{C}^M$ and more refined

$$\mathbf{z}_l = \begin{bmatrix} \mathbf{z}_l[1] \\ \vdots \\ \mathbf{z}_l[Q'] \end{bmatrix},$$

where $\mathbf{z}_l[i] \in \mathbb{C}^d$.

Since $\mathbf{z} \in \ker(\mathbf{D})$, for every l we have $\mathbf{B}_l\mathbf{z}_l = -\sum_{l'\neq l} \mathbf{B}_{l'}\mathbf{z}_{l'}$. We multiply this from the left with $\mathbf{B}_l^H$ to get

$$\mathbf{z}_l = -\sum_{l'\neq l} \mathbf{B}_l^H \mathbf{B}_{l'}\mathbf{z}_{l'}.$$

With $\mathbf{A}_{l,l'} := \mathbf{B}_l^H\mathbf{B}_{l'}$ the summands on the right side of the above equation are of the form

$$\mathbf{B}_l^H\mathbf{B}_{l'}\mathbf{z}_{l'} = \begin{bmatrix} \mathbf{A}_{l,l'}[1,1]\,\mathbf{z}_{l'}[1] + \dots + \mathbf{A}_{l,l'}[1,Q']\,\mathbf{z}_{l'}[Q'] \\ \mathbf{A}_{l,l'}[2,1]\,\mathbf{z}_{l'}[1] + \dots + \mathbf{A}_{l,l'}[2,Q']\,\mathbf{z}_{l'}[Q'] \\ \vdots \\ \mathbf{A}_{l,l'}[Q',1]\,\mathbf{z}_{l'}[1] + \dots + \mathbf{A}_{l,l'}[Q',Q']\,\mathbf{z}_{l'}[Q'] \end{bmatrix}, \tag{3.9}$$

where $\mathbf{A}_{l,l'}[i,j] = \mathbf{B}_l[i]^H\,\mathbf{B}_{l'}[j]$. Hence all $\mathbf{A}_{l,l'}[i,j]$ obey $\rho\big(\mathbf{A}_{l,l'}[i,j]\big) \le d\,\mu_B$ due to the definition of the block-coherence. That is, their spectral norm is bounded above by $d\,\mu_B$.

We compute the ℓ_2-norm of the d-row blocks of (3.9) and make us of the sub-multiplicity and the triangle inequality, of the matrix norm, as well as $\rho\big(\mathbf{A}_{l,l'}[i,j]\big) \le d\,\mu_B$ to get the elementwise inequality

$$\begin{aligned}\begin{bmatrix} \|\mathbf{z}_l[1]\|_2 \\ \vdots \\ \|\mathbf{z}_l[Q']\|_2 \end{bmatrix} &\le \sum_{l' \neq l} \begin{bmatrix} d\,\mu_B\, I(\|\mathbf{z}_l[1]\|_2)\,(\|\mathbf{z}_{l'}[1]\|_2 + \ldots + \|\mathbf{z}_{l'}[Q']\|_2) \\ \vdots \\ d\,\mu_B\, I(\|\mathbf{z}_l[Q']\|_2)\,(\|\mathbf{z}_{l'}[1]\|_2 + \ldots + \|\mathbf{z}_{l'}[Q']\|_2) \end{bmatrix} \\ &= d\,\mu_B \sum_{l' \neq l} \begin{bmatrix} I(\|\mathbf{z}_l[1]\|_2)\,\|\mathbf{z}_{l'}\|_{2,1} \\ \vdots \\ I(\|\mathbf{z}_l[Q']\|_2)\,\|\mathbf{z}_{l'}\|_{2,1} \end{bmatrix}.\end{aligned}$$

Note that we introduced the term $I(\|\mathbf{z}_l[j]\|_2)$ on the right hand side of the inequality to obtain a better estimate in the sequel (if an element of the left hand side of the inequality is zero, we can also set the corresponding element on the right hand side to zero).

Summing over all entries of the above vectors yields

$$\begin{aligned}\|\mathbf{z}_l\|_{2,1} &\le d\,\mu_B \sum_{j=1}^{Q'} \sum_{l' \neq l} I(\|\mathbf{z}_l[j]\|_2)\,\|\mathbf{z}_{l'}\|_{2,1} \\ &= d\,\mu_B \left(\sum_{j=1}^{Q'} I(\|\mathbf{z}_l[j]\|_2)\right)\left(\sum_{l' \neq l} \|\mathbf{z}_{l'}\|_{2,1}\right) \\ &= d\,\mu_B\,\|\mathbf{z}_l\|_{2,0}\left(\sum_{l' \neq l} \|\mathbf{z}_{l'}\|_{2,1}\right).\end{aligned}$$

We add $d\,\mu_B\,\|\mathbf{z}_l\|_{2,0}\|\mathbf{z}_l\|_{2,1}$ to make the sum on the right run over all l, that is

$$(1 + d\,\mu_B\,\|\mathbf{z}_l\|_{2,0})\,\|\mathbf{z}_l\|_{2,1} \le d\,\mu_B\,\|\mathbf{z}_l\|_{2,0} \sum_{l'} \|\mathbf{z}_{l'}\|_{2,1},$$

or equivalently

$$\begin{aligned}\|\mathbf{z}_l\|_{2,1} &\le \frac{d\,\mu_B\,\|\mathbf{z}_l\|_{2,0}}{1 + d\,\mu_B\,\|\mathbf{z}_l\|_{2,0}} \sum_{l'} \|\mathbf{z}_{l'}\|_{2,1} \\ &= \frac{d\,\mu_B\,\|\mathbf{z}_l\|_{2,0}}{1 + d\,\mu_B\,\|\mathbf{z}_l\|_{2,0}} \|\mathbf{z}\|_{2,1}.\end{aligned}$$

Then we sum over all l to get

$$1 \le \sum_{l=1}^{L} \frac{d\,\mu_B\,\|\mathbf{z}_l\|_{2,0}}{1 + d\,\mu_B\,\|\mathbf{z}_l\|_{2,0}}.$$

By adding $\sum_{l=1}^{L} \frac{1}{1+d\,\mu_B\,\|\mathbf{z}_l\|_{2,0}}$ to both sides, we obtain

$$\sum_{l=1}^{L} \frac{1}{1+d\,\mu_B\,\|\mathbf{z}_l\|_{2,0}} \leq L-1.$$

Note that the function $f(\mathbf{a}) := 1/(1+d\,\mu_B\,\|\mathbf{a}\|_{2,0})$ is convex in $\mathbf{a}$. Thus $[\sum_{l=1}^{L} f(\mathbf{z}_l)]/L \geq f(\sum_{l=1}^{L} \mathbf{z}_l/L)$ by Jensens' inequality [24]. Therefore,

$$\frac{1}{1+d\,\mu_B\,\|\mathbf{z}\|_{2,0}/L} \leq \frac{L-1}{L},$$

and finally

$$\|\mathbf{z}\|_{2,0} \geq \left(1+\frac{1}{L-1}\right)\frac{1}{d\,\mu_B}.$$

□

Proof of Proposition 3.1. Apply Lemma 3.2 to Lemma 3.3. □

3.1.2 Theoretical Discussion

In this section we want to emphasize the possible advantage of the derived condition in Proposition 3.1 over similar conditions for general dictionaries and corresponding results for the sparse case. For this assume $\mathbf{D}$ to be a dictionary consisting of L orthonormal bases with coherence μ and block-coherence μ_B.

For ease of reading we restate the conditions that will be considered in this discussion. A sufficient condition on k to guarantee uniqueness of the block-sparse representation problem for all $\mathbf{y} \in \mathcal{Y}_k$ (derived without taking the special structure of $\mathbf{D}$ into account) is

$$k < \frac{1}{2}\left(1+\frac{1}{d\,\mu_B}\right), \tag{3.10}$$

while the condition derived in this thesis under the assumption of a dictionary consisting of L orthonormal bases is

$$k < \frac{1}{2}\left(1+\frac{1}{L-1}\right)\frac{1}{d\,\mu_B}. \tag{3.11}$$

The corresponding equations for the regular sparse case are

$$k' < \frac{1}{2}\left(1+\frac{1}{\mu}\right), \tag{3.12}$$

and

$$k' < \frac{1}{2}\left(1+\frac{1}{L-1}\right)\frac{1}{\mu}, \tag{3.13}$$

respectively, where a k' satisfying the condition ensures that the regular sparse representation problem has a unique solution—note that this guarantees the uniqueness of the block-sparse representation problem for $k = \lfloor k'/d \rfloor$.

We are always only interested in the largest k and k' allowed by the mentioned conditions. Hence, when we refer to the block-sparsity (sparsity) allowed by some condition we mean the largest k (k') obeying this condition.

When we say that a condition on the block-sparsity allows for a higher sparsity than some condition on the regular sparsity, we actually mean that the block-sparsity allowed by the former condition multiplied by d is larger than the sparsity allowed by the latter one (kd equals the number of coefficients in k blocks of block-size d). We also say that the sparsity is potentially higher in this case.

Comparison to the Sparse Case

For a comparison of the previous mentioned conditions for the sparse case and the block-sparse case we have to relate the coherence μ and the block-coherence μ_B of a dictionary that consists of the union of L orthonormal bases. By Lemma C.1, we have

$$\frac{\mu}{d} \leq \mu_B \leq \mu. \tag{3.14}$$

Condition (3.11) can be an improvement over the corresponding condition for the sparse case (3.13). It allows for a sparsity of up to

$$k' \leq d\,k < \frac{1}{2}\left(1 + \frac{1}{L-1}\right)\frac{1}{\mu_B}. \tag{3.15}$$

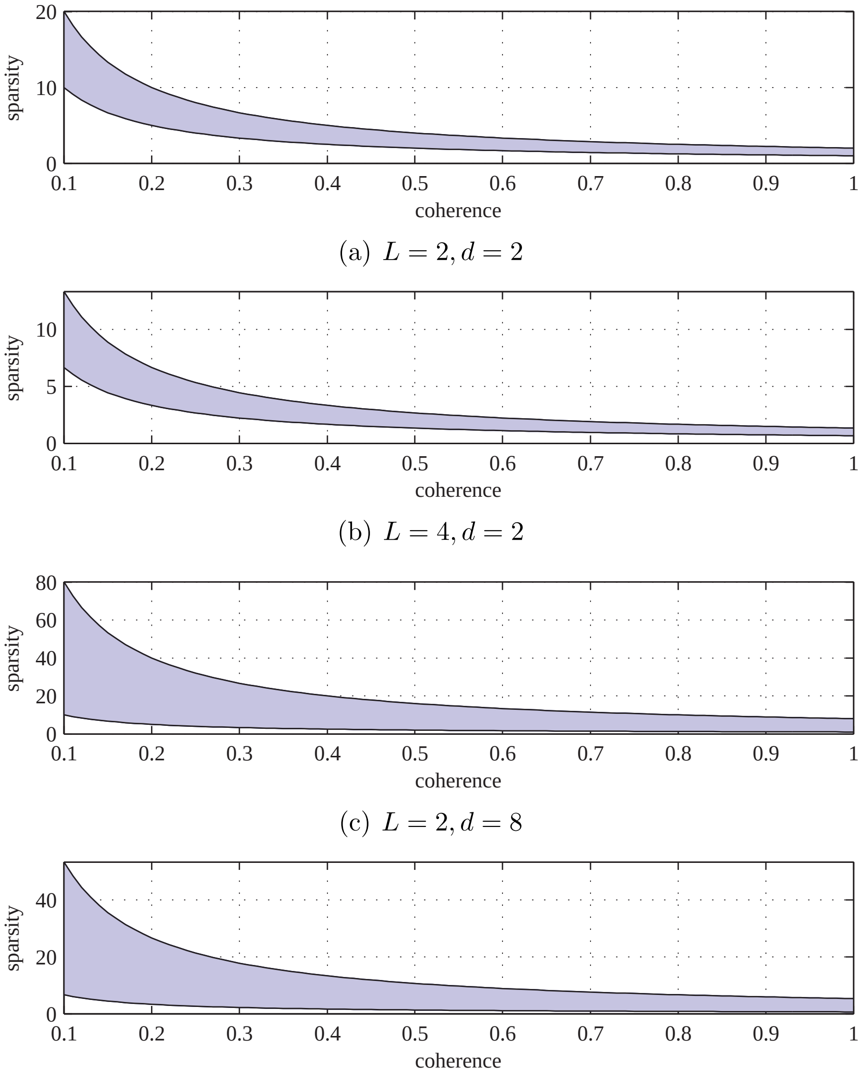
By (3.14) we have $\mu_B \leq \mu$. Therefore, Condition (3.11) allows for a potentially higher sparsity than Condition (3.13).

From (3.14) we also have $\mu/d \leq \mu_B$ as a lower bound on the block-coherence. Hence, Condition (3.11) can allow for a sparsity that is at most larger by a factor of the block-size d over the sparsity allowed by the corresponding condition for the sparse case.

The possible advantage of Condition (3.11) over Condition (3.13) for the sparse case is illustrated in Figure 3.1.

Comparison to the Uniqueness Condition for General Dictionaries

We fix a dictionary that is comprised of the union of L orthonormal bases. Comparing the block-sparsity allowed by the Condition (3.11) for dictionaries that are unions of orthonormal bases to the block-sparsity allowed by the Condition (3.10) for general dictionaries, reveals that the former condition allows for a larger block-sparsity, whenever $L < [1/(d\,\mu_B) + 1]$. Since the considered dictionary has orthonormal blocks, the block-coherence obeys $\mu_B \leq 1/d$ according to [16], Proposition 3. Thus, if $L = 2$, then Condition (3.11) allows for at least the same



(a) $L = 2, d = 2$

(b) $L = 4, d = 2$

(c) $L = 2, d = 8$

(d) $L = 4, d = 8$

Figure 3.1: Possible advantage of Condition (3.11) over the corresponding Condition (3.13) for the sparse case. The lower curve in the plot shows the sparsity allowed by (3.13) against the coherence. The potential sparsity of a dictionary with a certain coherence μ' allowed by Condition (3.11) can lie somewhere in the blue region on the vertical line going through μ' on the x-axis (the position in the blue region on the line depends on the real block-coherence μ_B of the dictionary). Each plot shows a different combination of the block-size d and the number of dictionaries L. Not all pairs of μ and μ_B may be possible.

block-sparsity as the condition for general dictionaries for all possible values of the block-coherence.

Uniqueness Versus Richness

In general, we have to make a trade-off between the number of orthonormal bases L in a dictionary and the maximal block-sparsity k that guarantees uniqueness of the block-sparse representation problem. While with an increasing number of orthonormal bases we usually can make the representation of signals more efficient, i.e., we need to use less nonzero blocks to represent a signal, due to the greater number of different blocks available, the block-sparsity allowed by Condition (3.11) typically decreases due to an increase of L and the block-coherence μ_B.

Figure 3.2 shows how the block-sparsity allowed by (3.11) and the number of blocks in the dictionary develop with the number of orthonormal bases L. The left y-axis is normalized by multiplication with $d\,\mu_B$ and the right y-axis by division over the number of blocks per dictionary.

We observe that for classes of signals that require a large number of bases to have an efficient representation in the dictionary, it may be a good choice to use a large number of bases in the dictionary. In this way it may be possible to decrease the block-sparsity necessary for the representation of the considered signals in the dictionary while not paying a big price on the sparsity allowed by (3.11).

For low values of L the trade-off should be evaluated case-by-case.

Note that depending on the values of L and of the block-coherence μ_B, the condition for general dictionaries may be better.

Furthermore, it may not be possible to create a dictionary with some specific L and μ_B.

Dictionary Structure

Another interesting fact is that for some fixed number of orthonormal bases $L \geq 3$ in the dictionary there is a point regarding the block-sparsity, starting from which Condition (3.11) gets worse than the condition for general dictionaries (3.10). This is illustrated in Figure 3.3. The crossover point up to which Condition (3.11) is better than Condition (3.10) shifts to the left with an increasing number of orthonormal bases L.

3.1.3 Simulation Results

The aim of this section is to demonstrate some of the results from the theoretical discussion.

For this we created dictionaries consisting of the union of two orthonormal bases ranging over a subset of the possible values of the block-coherence μ_B, namely

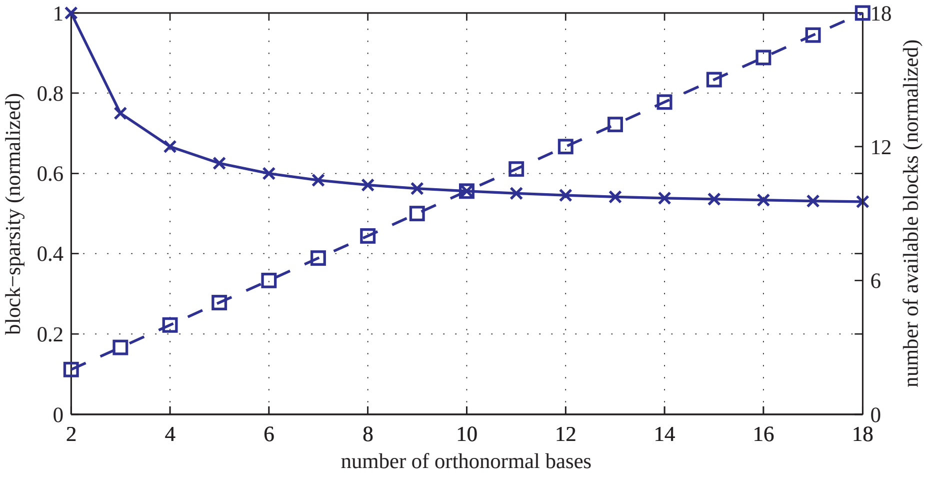


Figure 3.2: Change of the block-sparsity allowed by (3.11) (solid line) and the number of available blocks for signal representation (dashed line) against the number of orthonormal bases L in the dictionary. The allowed block-sparsity is normalized by multiplication with $d\,\mu_B$ and the number of available blocks is normalized by division over the number of blocks per basis.

dictionaries with a block-coherence that satisfies

$$\frac{1}{\sqrt{Md}} \leq \mu_B \leq \frac{1}{d}, \tag{3.16}$$

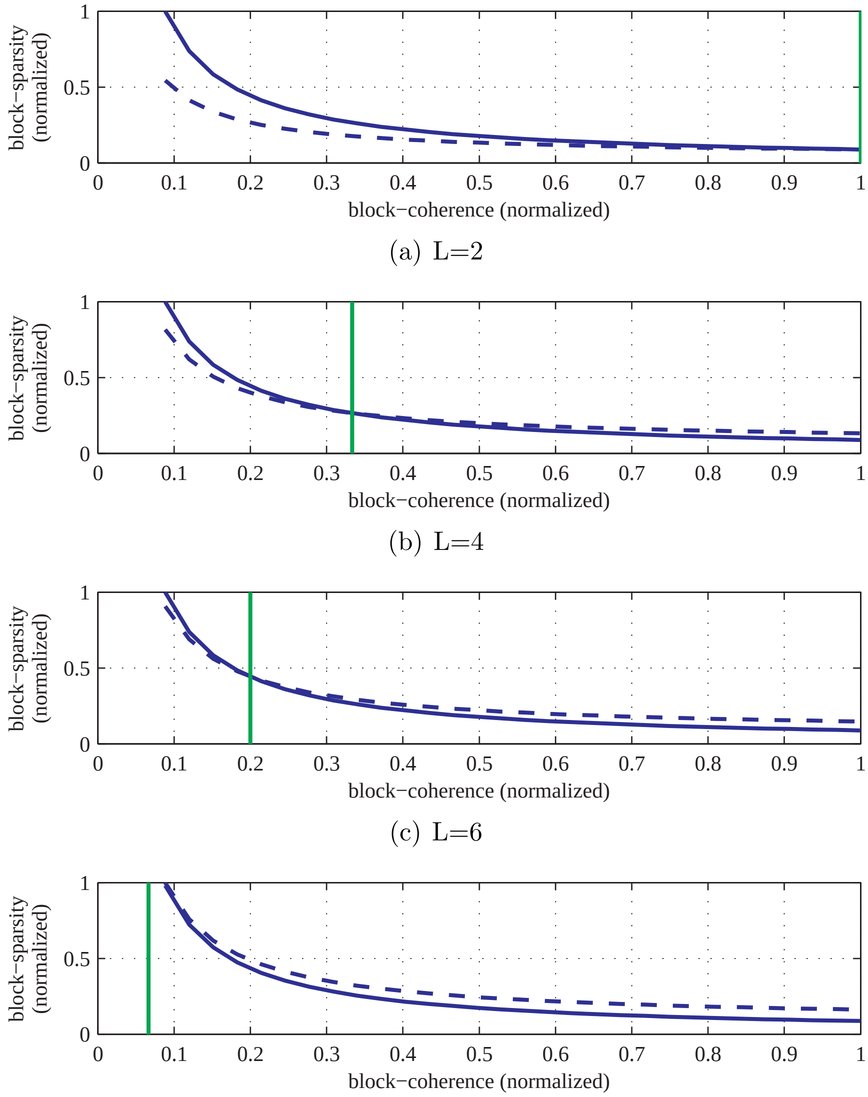
where M is the dimension of the space spanned by the dictionary—the upper and lower bound on the block-coherence μ_B are from [16].

For the created dictionaries we calculated the block-coherence μ_B, as well as the coherence μ and compared the allowed sparsity by (3.13) and (3.12) to the potential sparsity allowed by (3.10) and (3.11). For the created dictionaries we varied the dimension M, as well as the block-size d.

Figure 3.4 shows two such comparisons. The upper two curves in the figure are smooth because they depend on the block-coherence μ_B which we vary on the x-axis, while the lower two curves depend on the coherence μ which is *arbitrary* for the dictionaries we created (except for the rightmost point which corresponds to a dictionary consisting of two times the same basis and thus the coherence is clearly 1).

The potential sparsities allowed by the conditions for the block-sparse case (upper two curves) are significantly higher than the allowed sparsity by the conditions for the general sparse case (lower two curves). Thus, taking the model-assumption of block-sparsity into account can be advantageous.

Further, note that the potential sparsity allowed by the condition for dictionaries consisting of the union of orthonormal bases lies above the potential sparsity allowed by the condition for general dictionaries in the block-sparse case. At the



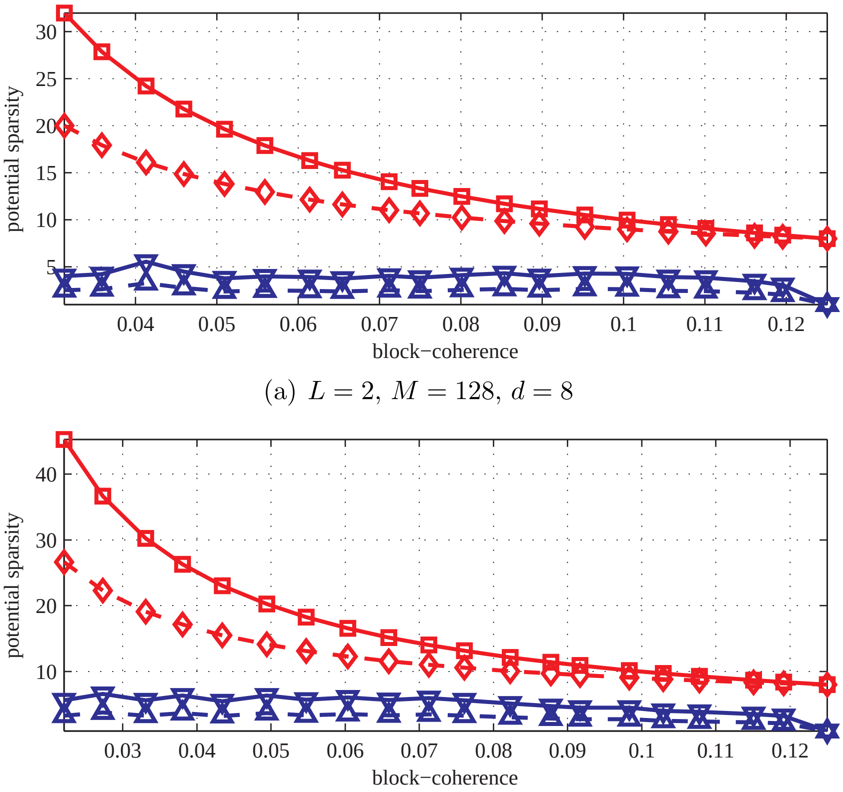
(a) L=2

(b) L=4

(c) L=6

(d) L=16

Figure 3.3: Block-sparsity allowed by (3.10) (dashed line) and (3.11) (solid line) against the block-coherence multiplied by the block-size d. The plots compare the two conditions for dictionaries consisting of different numbers L of orthonormal bases. The vertical solid line marks the crossover point starting from which Condition (3.10) allows for a higher block-sparsity when increasing the block-coherence. Note that for dictionaries consisting of orthonormal blocks the block-coherence obeys $\mu_B \leq 1/d$ and hence the x-axis goes only up to 1.



(a) $L = 2$, $M = 128$, $d = 8$

(b) $L = 2$, $M = 256$, $d = 8$

Figure 3.4: Comparison of the potential sparsity allowed by (3.10) (red dashed line), (3.11) (red solid line) and the sparsity allowed by (3.12) (blue dashed line) and (3.13) (blue solid line) for dictionaries comprised of $L = 2$ orthonormal bases, varying dimension M, and with block-size $d = 8$.

right border of the figure, the two curves for the block-sparsity touch each other. This point in the figure corresponds to a dictionary with block-coherence $1/d$, for which both conditions give the same limit on the block-sparsity for $L = 2$.

3.2 Signal Recovery through ℓ_2/ℓ_1 Optimization

In this section we consider the problem of recovering the representation $\mathbf{x}$ of a vector $\mathbf{y}$ in a dictionary $\mathbf{D}$ that is comprised of L orthonormal bases, i.e., we want to find the block-sparsest $\mathbf{x}$ such that $\mathbf{y} = \mathbf{D}\mathbf{x}$ from $\mathbf{y}$ and $\mathbf{D}$.

As mentioned in Chapter 2, the representation $\mathbf{x}$ is not necessarily unique for a fixed $\mathbf{y}$. That is, there can be vectors $\mathbf{x}' \neq \mathbf{x}$ such that $\mathbf{y} = \mathbf{D}\mathbf{x}'$. However, by requiring $\mathbf{y} \in \mathcal{Y}_k$, where k is small enough, we can guarantee that there is a unique sparsest representation for all possible $\mathbf{y}$.

If k is sufficiently small, we can furthermore ensure that the block-sparsest representation of any $\mathbf{y} \in \mathcal{Y}_k$ satisfies the Exact Recovery Condition in Lemma 2.2. Hence, in this case the block-sparsest representation of any possible $\mathbf{y}$ can be recovered using L-OPT and BOMP.

Our contribution, which is summarized in Proposition 3.2, is a sufficient condition on k such that for all $\mathbf{y} \in \mathcal{Y}_k$ the corresponding block-sparsest representation can be recovered by L-OPT and BOMP. We demonstrate by a theoretical discussion and simulation results that our sufficient condition can allow for a higher block-sparsity than the condition in Proposition 2.6. Further, we show that our condition is advantageous over the similar conditions (2.18) and (2.20) for the regular sparse case if the block-sparsity assumption is applicable, as it guarantees that representations with a higher sparsity can be recovered.

3.2.1 Main Results

Assume that $\mathbf{y} = \mathbf{D}\mathbf{x}$. As before, we denote by $\mathbf{D}_{opt}$ a matrix comprised of the blocks of $\mathbf{D}$ that correspond to the nonzero blocks of $\mathbf{x}$ and by $\overline{\mathbf{D}}_{opt}$ a matrix that consists of the blocks of $\mathbf{D}$ that are not in $\mathbf{D}_{opt}$. Then the Exact Recovery Condition for the block-sparse case, i.e., the condition

$$\rho_c\Big(\mathbf{D}_{opt}^{\dagger}\overline{\mathbf{D}}_{opt}\Big) < 1,$$

is sufficient to guarantee that L-OPT and BOMP recover the representation $\mathbf{x}$ from $\mathbf{y}$ exactly (see Section 2.2.5 for details) [16].

By following the path taken in [19], we derived the following proposition.

Proposition 3.2. *Let the dictionary* $\mathbf{D}$ *consist of* L *orthonormal bases with block-size* d *and let* $\mathbf{y} \in \mathcal{Y}_k$. *If*

$$k < \left(\sqrt{2} - 1 + \frac{1}{2\,(L-1)}\right)\frac{1}{d\,\mu_B}, \tag{3.17}$$

then the block-sparsest representation of $\mathbf{y}$ *can be recovered by L-OPT and BOMP.*

Furthermore, if $\mathbf{y} = \mathbf{D}\mathbf{x}$, *where* $\|\mathbf{x}\|_{2,0} \leq k$, *then* $\mathbf{x}$ *is the unique block-sparsest representation of* $\mathbf{y}$.

Proof. See Appendix D. □

3.2.2 Theoretical Discussion

In this section we consider under which conditions and up to which extend the condition given by Proposition 3.2 can be advantageous over other known results. For this we assume the dictionary $\mathbf{D}$ to consist of the union of L orthonormal bases.

When we compare conditions and say that some condition is better than another condition, we mean that the former condition allows for a higher block-sparsity (sparsity) than the latter condition.

Advantage Over Sparse Case

We compare Condition (3.17) with the corresponding condition for the regular sparse case (see Section 2.1.5 for details), i.e., with

$$k' < \left(\sqrt{2} - 1 + \frac{1}{2\,(L-1)}\right)\frac{1}{\mu}, \tag{3.18}$$

By Condition (3.18) a representation $\mathbf{x}$ such that $\mathbf{y} = \mathbf{D}\mathbf{x}$ can be recovered from $\mathbf{y}$ and $\mathbf{D}$ by BP and OMP if $\|\mathbf{x}\|_0 \leq k'$. Note that (3.17) guarantees recovery with L-OPT and BOMP for a sparsity of up to

$$d\,\|\mathbf{x}\|_{2,0} < \left(\sqrt{2} - 1 + \frac{1}{2\,(L-1)}\right)\frac{1}{\mu_B}. \tag{3.19}$$

Since $\mu_B \leq \mu$ it can be possible to recover signals with a higher sparsity by L-OPT and BOMP than by BP and OMP if the block-sparsity assumption is applicable.

Example 3.1. *Consider the dictionary* $\mathbf{D} = [\mathbf{I}, \mathbf{F} \otimes \mathbf{I}_d] \in \mathbb{C}^{M \times 2M}$*, where* $\mathbf{F}$ *denotes the DFT-matrix of size* $(M/d) \times (M/d)$ *and* $\otimes$ *the Kronecker product. This dictionary has block-coherence* $\mu_B = 1/\sqrt{Md}$ *and coherence* $\mu = 1/\sqrt{M}$*. Hence,* (3.17) *allows for a sparsity that is larger by a factor of* $\sqrt{d}$ *than* (3.18)*.*

Advantage over Recovery Conditions for Arbitrary Dictionaries

Some similar comments as in this section were made in [14].

Comparing the Condition (3.17) to the condition for arbitrary dictionaries in the block-sparse case (see Section C.3 in the Appendix), i.e., to

$$k < \frac{1}{2}\left(1 + \frac{1}{d\,\mu_B}\right), \tag{3.20}$$

reveals that our condition is not better for all numbers of dictionaries L.

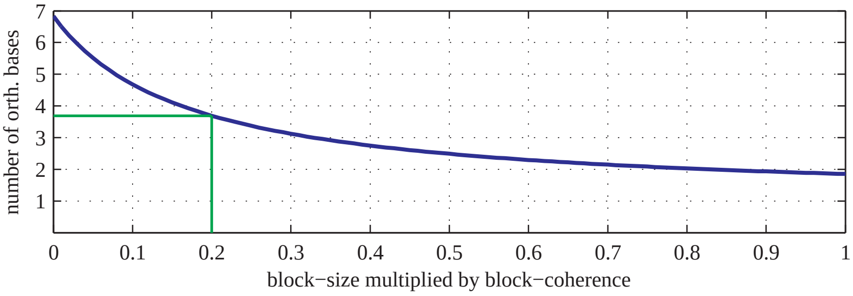


Figure 3.5: Comparison of Condition (3.17) and (3.20). For a dictionary consisting of L orthonormal bases with block-size d and block-coherence μ_B the former condition, i.e., the condition derived for dictionaries consisting of orthonormal bases, allows for a higher block-sparsity if the number of orthonormal bases L in the dictionary is below the value corresponding to the intersection of a vertical line drawn at $d\,\mu_B$ with the curve. For example, for a dictionary with block-coherence μ_B and block-size d such that $d\,\mu_B = 0.2$ Condition (3.17) is better if the number of orthonormal bases in the dictionary is at most 3 (green lines in the plot).

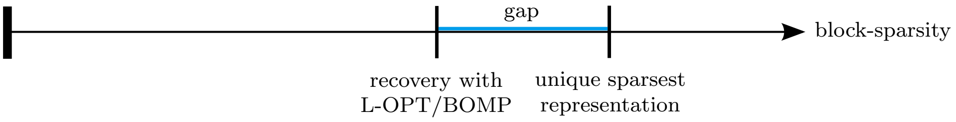


Figure 3.6: Schematic sketch of the gap between the maximal block-sparsity allowed by Condition (3.17) and Condition (3.11).

It is only better, whenever

$$L < \frac{1}{3 - 2\sqrt{2} + d\,\mu_B} + 1. \tag{3.21}$$

Since $d\,\mu_B$ is bounded below by 0, Condition (3.17) can only be better than Condition (3.20) for $L \leq 6$. This is illustrated in Figure 3.5.

Observe that there is a gap between the block-sparsity allowed by Condition (3.17) and Condition (3.11)—which both were derived taking the special structure of a dictionary consisting of the union of L orthonormal bases into account. This means that there is a region of the block-sparsity k for which any block k-sparse vector $\mathbf{x}$ is the unique block-sparsest representation of some signal $\mathbf{y} = \mathbf{D}\mathbf{x}$ but for which we have no guarantee that we can recover $\mathbf{x}$ by L-OPT and BOMP. This is schematically shown in Figure 3.6. However, note that other conditions, e.g., Condition (3.20), may guarantee successful recovery by L-OPT or BOMP .

3.2.3 Simulation Results

The presented simulations demonstrate some of the possible advantages discussed in the previous section.

For our simulations we created random matrices $\mathbf{A}_1, \mathbf{A}_2, \ldots$ of size $M \times LM$ with entries drawn from an i.i.d. Gaussian distribution. L is the number of orthonormal bases and $M = Qd$ their dimension, where M, Q, d are integers and d is the block-size. From each of these matrices $\mathbf{A}_i$ we created a random dictionary $\mathbf{D}_i$ as follows: For $l = 1, \ldots, L$ we set the lth M columns of $\mathbf{D}_i$ to the orthonormal basis created from the lth M columns of $\mathbf{A}_i$ by the MATLAB command `orth`.

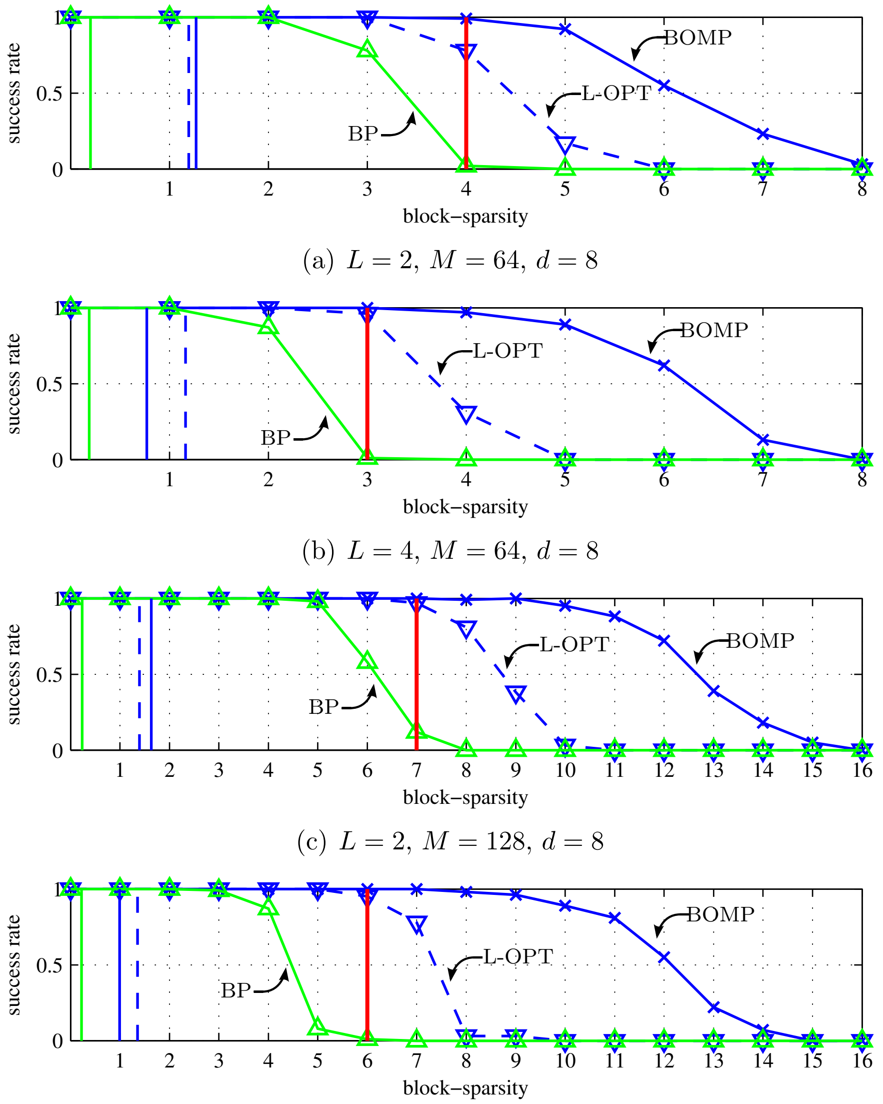
For these dictionaries we compared the recovery *success rate* of L-OPT, BOMP and BP. That is, we created block k-sparse vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots$ of length LM, where the k nonzero blocks were selected uniformly at random and the entries within the blocks are i.i.d. Gaussian. For these vectors we calculated $\mathbf{y}_i = \mathbf{D}_i\mathbf{x}_i$. Then, we provided $\mathbf{D}_i$ and $\mathbf{y}_i$ to L-OPT, BOMP and BP. All three algorithms produce an estimate $\mathbf{x}_{i,est}$ of the representation $\mathbf{x}_i$. The recovery was successful if $\|\mathbf{x}_i - \mathbf{x}_{i,est}\|_2 < \epsilon$, where $\epsilon = 10^{-6}$ is a small positive constant that allows for calculation inaccuracies. For each considered value of the block-sparsity k we averaged over 100 pairs of realizations of the dictionary and the block-sparse vector. The success rate is the number of successful recoveries out of 100.

The results of these simulations for a varying number of orthonormal bases L and dictionary dimensions M are shown in Figure 3.7. The block-size $d = 8$ is fixed. There are some observations to make:

- L-OPT clearly outperforms BP, which itself is outperformed by BOMP.
- For any of the plots: Let k' be the minimal block-sparsity for which L-OPT failed at least once in recovering the underlying representation and let k be the maximum block-sparsity allowed by Condition (3.17) or (3.20) (rounded down to the nearest integer). Then k' is at least larger by two than k. That is, we observed a 100 percent success rate in recovering the representations with a block-sparsity of up to $k' - 1$ by L-OPT. This implies that
  - the highest possible block-sparsity k such that the block-sparsest representation of any $\mathbf{y} \in \mathcal{Y}_k$ can be recovered by L-OPT is larger than the block-sparsity allowed by the considered conditions, or
  - that the equivalence of the ℓ_2/ℓ_0 and ℓ_2/ℓ_1 minimization problems holds with high probability for the generated random signals even if their block-sparsity is between k and k'.

  Also a combination of the above implications is possible.

Maximum Block-Incoherent Dictionary. We repeated the previous experiment but instead of the random dictionaries used the fixed dictionary $\mathbf{D} =$



(a) $L = 2$, $M = 64$, $d = 8$

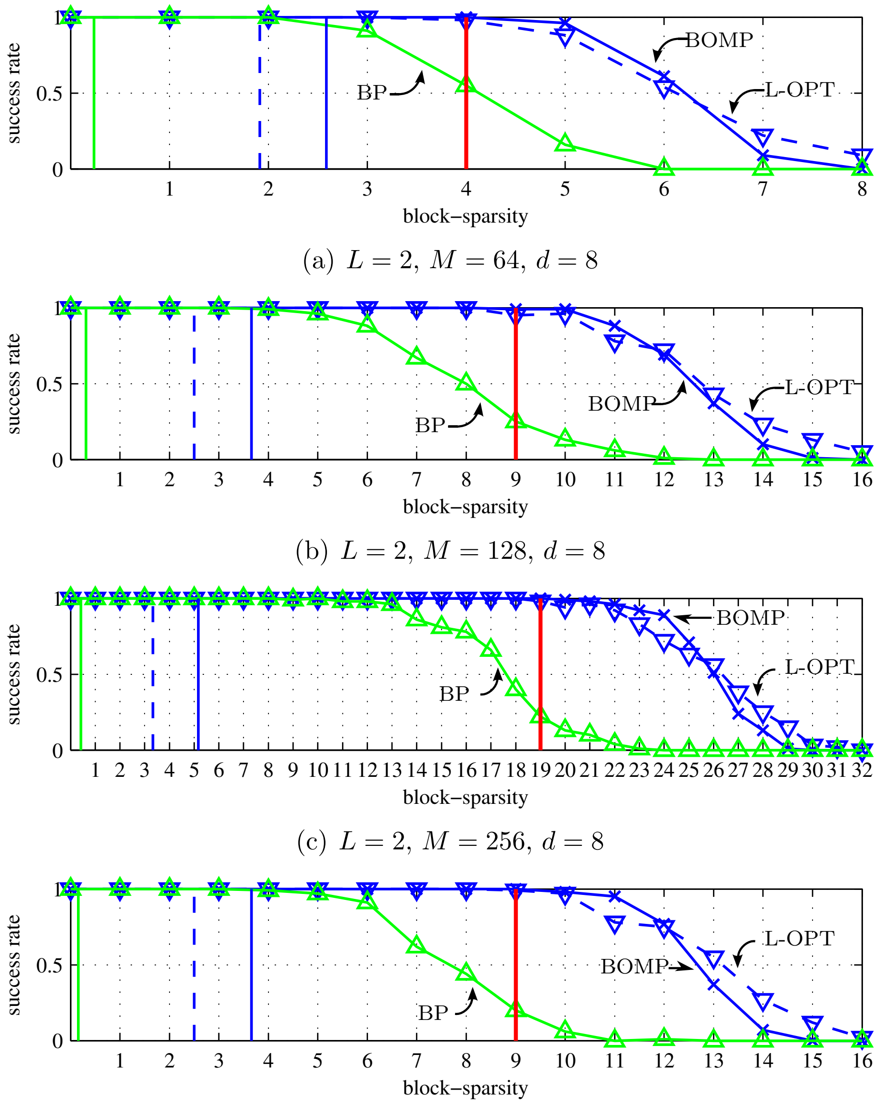
(b) $L = 4$, $M = 64$, $d = 8$

(c) $L = 2$, $M = 128$, $d = 8$

(d) $L = 4$, $M = 128$, $d = 8$

Figure 3.7: Comparison of the recovery success rate of BP, L-OPT and BOMP in random dictionaries with block-size d consisting of the union of L orthonormal bases against the block-sparsity of the representation to recover. The vertical solid green line is the maximal sparsity allowed by Condition (3.18) (guaranteed recovery with BP in arbitrary dictionaries), the vertical dashed blue line shows the maximal block-sparsity allowed by Condition (3.20) (guaranteed recovery with BOMP and L-OPT in arbitrary dictionaries) and the vertical solid blue line is the maximal block-sparsity allowed by Condition (3.17) (guaranteed recovery with BOMP and L-OPT in dictionaries that are the unions of orthonormal bases)—these lines were obtained through averaging. The vertical red line is the minimal block-sparsity for which L-OPT failed at least once.

$[\mathbf{I}_M, \mathbf{F}_{M/d} \otimes \mathbf{I}_d]$. In this case $L = 2$. This dictionary achieves the lowest possible block-coherence for a dictionary consisting of two orthonormal bases, i.e., it has block-coherence $\mu_B = 1/\sqrt{M\, d}$ (see [16]). For every considered value of the block-sparsity we averaged over 100 realizations of the random block-sparse vector. The results of this experiment are shown in figure 3.8. In contrast to before L-OPT sometimes outperforms BOMP.



(a) $L = 2$, $M = 64$, $d = 8$

(b) $L = 2$, $M = 128$, $d = 8$

(c) $L = 2$, $M = 256$, $d = 8$

(d) $L = 2$, $M = 256$, $d = 16$

Figure 3.8: Comparison of the recovery success rate of BP, L-OPT and BOMP in the dictionary $\mathbf{D} = [\mathbf{I}_M, \mathbf{F}_{M/d} \otimes \mathbf{I}_d]$ with block-size d against the block-sparsity of the representation to recover. The vertical solid green line is the maximal sparsity allowed by Condition (3.18) (guaranteed recovery with BP in arbitrary dictionaries), the vertical dashed blue line shows the maximal block-sparsity allowed by Condition (3.20) (guaranteed recovery with BOMP and L-OPT in arbitrary dictionaries) and the vertical solid blue line is the maximal block-sparsity allowed by Condition (3.17) (guaranteed recovery with BOMP and L-OPT in dictionaries that are the unions of orthonormal bases)—these lines were obtained through averaging. The vertical red line is the minimal block-sparsity for which L-OPT failed at least once.

Chapter 4

Block-Sparse Approximation

Consider the problem of representing an arbitrary signal $\mathbf{y} \in \mathbb{C}^M$ in a fixed dictionary $\mathbf{D}$ by a block k-sparse vector $\mathbf{x} \in \mathbb{C}^N$ such that $\mathbf{y} = \mathbf{D}\mathbf{x}$. This is in general not possible because the atoms within any selection of k blocks from the dictionary do not necessarily span the space $\mathbb{C}^M$.

We are then interested in finding a k-block approximation $\mathbf{a}$ of $\mathbf{y}$. The quality of such an approximation is measured by an *error criterion*. We use the ℓ_2-norm of $\mathbf{y} - \mathbf{a}$, i.e., $\|\mathbf{y} - \mathbf{a}\|_2$, as the error criterion because it gives rise to simple algebraic results. The lower the error criterion, the better is the approximation.

4.1 BOMP as an Approximation Algorithm

In this section we investigate the performance of BOMP as an approximation algorithm. That is, how well the algorithm solves the block-sparse approximation problem described in Section 2.2.2.

We briefly restate the problem: For an arbitrary signal $\mathbf{y} \in \mathbb{C}^M$, a dictionary $\mathbf{D}$ of size $M \times N$ with block-size d, find a block k-sparse vector $\mathbf{x}$ as the solution to the problem

$$\arg \min_{\mathbf{x} \in \mathbb{C}^N} \|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2 \quad \text{s.t.} \|\mathbf{x}\|_{2,0} \leq k. \tag{4.1}$$

We associate the approximation $\mathbf{a}(\mathbf{x}') = \mathbf{D}\mathbf{x}'$ with $\mathbf{x}'$. An approximation $\mathbf{a}(\mathbf{x}')$ is *optimal* if $\mathbf{x}'$ is a solution to (4.1). The *approximation error* of $\mathbf{a}$ for $\mathbf{y}$ is $\|\mathbf{y} - \mathbf{a}\|_2$. Whenever it is clear which $\mathbf{x}'$ we refer to, we write $\mathbf{a}$ instead of $\mathbf{a}(\mathbf{x}')$. Note that the above minimization problem can in general have many solutions. This is demonstrated by the following example.

Example 4.1 (Optimal k-Block Approximation)**.** *Let $d = 2$ be the block-size.*

Consider the dictionary

$$\mathbf{D} = \left[\begin{array}{cc|cc|cc} 1 & 0 & 0 & 0 & 0 & 1/\sqrt{5} \\ 0 & 1 & 0 & 0 & 0 & 1/\sqrt{5} \\ 0 & 0 & 1 & 0 & 0 & 1/\sqrt{5} \\ 0 & 0 & 0 & 1 & 0 & 1/\sqrt{5} \\ 0 & 0 & 0 & 0 & 1 & 1/\sqrt{5} \end{array}\right],$$

and the signal

$$\mathbf{y} = [0, 1, 1, 0, 0]^T .$$

The optimal 1*-block approximations for* $\mathbf{y}$ *in* $\mathbf{D}$ *are*

$$\begin{aligned} &\mathbf{a}_{opt_1}(\mathbf{x}_1) = [0, 1, 0, 0, 0]^T \ \textit{with}\ \mathbf{x}_1 = [0, 1, 0, 0, 0, 0]^T , \\ &\mathbf{a}_{opt_2}(\mathbf{x}_2) = [0, 0, 1, 0, 0]^T \ \textit{with}\ \mathbf{x}_2 = [0, 0, 1, 0, 0, 0]^T , \textit{ and} \\ &\mathbf{a}_{opt_3}(\mathbf{x}_3) = [0.5, 0.5, 0.5, 0.5, 0]^T \ \textit{with}\ \mathbf{x}_3 = \left[0, 0, 0, 0, -0.5, \sqrt{5}/2\right]^T . \end{aligned}$$

Now fix an arbitrary signal $\mathbf{y}$. For this signal we obtain a k-block approximation $\mathbf{a}_k$ by BOMP in k-steps. To measure the approximation performance of BOMP we compare the approximation error of $\mathbf{a}_k$ for $\mathbf{y}$ to the approximation error of an optimal k-block approximation $\mathbf{a}_{opt}$ for the same signal $\mathbf{y}$. We relate the approximation errors by

$$\|\mathbf{y} - \mathbf{a}_k\|_2 \leq c \, \|\mathbf{y} - \mathbf{a}_{opt}\|_2,$$

where $c \geq 1$ is a real number. The smaller c is, the closer $\mathbf{a}_k$ is to an optimal k-block approximation and the better is the obtained approximation.

Our main result, that quantifies the quality of the approximation $\mathbf{a}_k$ with respect to the quality of an optimal approximation, is summarized in the following Proposition:

Proposition 4.1 (Adopted from [19], Corollary 4.3)**.** *Let* $\mathbf{D}$ *be a dictionary of size* $M \times N$ *with orthonormal blocks of block-size* d *and a block Babel function such that* $d\,\mu_{B1}(k) < 1/2$*, where* k *is a fixed positive integer, and let* $\mathbf{y} \in \mathbb{C}^M$ *be a completely arbitrary signal. BOMP will then recover a* k*-block approximation* $\mathbf{a}_k$ *of* $\mathbf{y}$ *in* k*-steps that satisfies*

$$\|\mathbf{y} - \mathbf{a}_k\|_2 \leq \sqrt{1 + \frac{k\,[1 - d\,\mu_{B1}(k)]}{[1 - 2\,d\,\mu_{B1}(k)]^2}}\, \|\mathbf{y} - \mathbf{a}_{opt}\|_2, \tag{4.2}$$

where $\mathbf{a}_{opt}$ *is an optimal* k*-block approximation of* $\mathbf{y}$*.*

4.1.1 Derivation

We start our analysis by deriving a condition similar to the Exact Recovery Condition given by Lemma 2.2 and later conclude how close we get to an optimal approximation when using BOMP as an approximation algorithm.

Lemma 4.1 (General Recovery, adopted from [19], Theorem 4.2)**.** *Let* $\mathbf{D}$ *be a dictionary of size* $M \times N$ *with orthonormal blocks of block-size* d *and a block Babel function such that* $d\,\mu_{B1}(k) < 1/2$*, where* k *is a fixed positive integer. Further, let* $\mathbf{y} \in \mathbb{C}^M$ *be an arbitrary signal. Suppose that the approximation* $\mathbf{a}_{j-1}$ *consists only of blocks from an optimal* k*-block approximation* $\mathbf{a}_{opt}$*. Then, at step* j*, BOMP will recover another block from* $\mathbf{a}_{opt}$*, whenever*

$$\|\mathbf{y} - \mathbf{a}_{j-1}\|_2 > \sqrt{1 + \frac{k\,[1 - d\,\mu_{B1}(k)]}{[1 - 2\,d\,\mu_{B1}(k)]^2}}\, \|\mathbf{y} - \mathbf{a}_{opt}\|_2. \tag{4.3}$$

Proof. The proof follows the arguments in [19].

We associate to $\mathbf{x}_{opt}$ the vector $\mathbf{a}_{opt} = \mathbf{D}\mathbf{x}_{\mathrm{opt}}$. Let $\mathbf{D}_{opt}$ be a matrix that consists of the blocks in $\mathbf{D}$ that correspond to the nonzero blocks of $\mathbf{x}_{opt}$. Furthermore, $\overline{\mathbf{D}}_{opt}$ consists of all blocks in $\mathbf{D}$ that do not participate in $\mathbf{D}_{opt}$.

According to the description of BOMP in Section 2.2.5, the algorithm selects another block from $\mathbf{D}_{opt}$ if

$$z(\mathbf{r}_{j-1}) := \frac{\|\overline{\mathbf{D}}_{opt}^H \mathbf{r}_{j-1}\|_{2,\infty}}{\|\mathbf{D}_{opt}^H \mathbf{r}_{j-1}\|_{2,\infty}} < 1. \tag{4.4}$$

We now introduce the quantities $\mathbf{a}_{j-1}$ and $\mathbf{a}_{opt}$ into the inequality above and bound the ratio. Note, that the mixed ℓ_2/ℓ_∞ norm is a matrix norm according to Lemma B.1. Thus,

$$\begin{aligned}
z(\mathbf{r}_{j-1}) &= \frac{\|\overline{\mathbf{D}}_{opt}^H \mathbf{r}_{j-1}\|_{2,\infty}}{\|\mathbf{D}_{opt}^H \mathbf{r}_{j-1}\|_{2,\infty}} \\
&= \frac{\|\overline{\mathbf{D}}_{opt}^H (\mathbf{y} - \mathbf{a}_{j-1})\|_{2,\infty}}{\|\mathbf{D}_{opt}^H (\mathbf{y} - \mathbf{a}_{j-1})\|_{2,\infty}} \\
&= \frac{\|\overline{\mathbf{D}}_{opt}^H (\mathbf{y} - \mathbf{a}_{opt}) + \overline{\mathbf{D}}_{opt}^H (\mathbf{a}_{opt} - \mathbf{a}_{j-1})\|_{2,\infty}}{\|\mathbf{D}_{opt}^H (\mathbf{y} - \mathbf{a}_{opt}) + \mathbf{D}_{opt}^H (\mathbf{a}_{opt} - \mathbf{a}_{j-1})\|_{2,\infty}} \\
&\leq \underbrace{\frac{\|\overline{\mathbf{D}}_{opt}^H (\mathbf{y} - \mathbf{a}_{opt})\|_{2,\infty}}{\|\mathbf{D}_{opt}^H (\mathbf{a}_{opt} - \mathbf{a}_{j-1})\|_{2,\infty}}}_{=: z_{err}(\mathbf{r}_{j-1})} + \underbrace{\frac{\|\overline{\mathbf{D}}_{opt}^H (\mathbf{a}_{opt} - \mathbf{a}_{j-1})\|_{2,\infty}}{\|\mathbf{D}_{opt}^H (\mathbf{a}_{opt} - \mathbf{a}_{j-1})\|_{2,\infty}}}_{=: z_{opt}(\mathbf{r}_{j-1})},
\end{aligned} \tag{4.5}$$

where $\mathbf{D}_{opt}^H(\mathbf{y} - \mathbf{a}_{opt})$ disappeared from the denominator because $\mathbf{y} - \mathbf{a}_{opt}$ is orthogonal to the subspace spanned by the columns of $\mathbf{D}_{opt}$.

We continue by bounding $z_{err}(\mathbf{r}_{j-1})$:

$$\begin{aligned}
z_{err}(\mathbf{r}_{j-1}) &= \frac{\|\overline{\mathbf{D}}_{opt}^{H}(\mathbf{y}-\mathbf{a}_{opt})\|_{2,\infty}}{\|\mathbf{D}_{opt}^{H}(\mathbf{a}_{opt}-\mathbf{a}_{j-1})\|_{2,\infty}} \\
&= \frac{\max_i \|(\overline{\mathbf{D}}_{opt}[i])^{H}(\mathbf{y}-\mathbf{a}_{opt})\|_{2}}{\|\mathbf{D}_{opt}^{H}(\mathbf{a}_{opt}-\mathbf{a}_{j-1})\|_{2,\infty}} \\
&\leq \frac{\max_i \|(\overline{\mathbf{D}}_{opt}[i])^{H}\|_2 \|\mathbf{y}-\mathbf{a}_{opt}\|_{2}}{\|\mathbf{D}_{opt}^{H}(\mathbf{a}_{opt}-\mathbf{a}_{j-1})\|_{2,\infty}} \\
&= \frac{\|\mathbf{y}-\mathbf{a}_{opt}\|_{2}}{\|\mathbf{D}_{opt}^{H}(\mathbf{a}_{opt}-\mathbf{a}_{j-1})\|_{2,\infty}},
\end{aligned}$$

because all blocks of $\mathbf{D}$ are orthonormal by assumption. Then

$$\begin{aligned}
z_{err}(\mathbf{r}_{j-1}) &\leq \frac{\|\mathbf{y}-\mathbf{a}_{opt}\|_{2}}{\|\mathbf{D}_{opt}^{H}(\mathbf{a}_{opt}-\mathbf{a}_{j-1})\|_{2,\infty}} \\
&\leq \frac{\sqrt{k}\,\|\mathbf{y}-\mathbf{a}_{opt}\|_{2}}{\|\mathbf{D}_{opt}^{H}(\mathbf{a}_{opt}-\mathbf{a}_{j-1})\|_{2}},
\end{aligned}$$

where the last inequality follows from Lemma B.3.

Continuing we find that

$$\begin{aligned}
z_{err}(\mathbf{r}_{j-1}) &\leq \frac{\sqrt{k}\,\|\mathbf{y}-\mathbf{a}_{opt}\|_{2}}{\|\mathbf{D}_{opt}^{H}(\mathbf{a}_{opt}-\mathbf{a}_{j-1})\|_{2}} \\
&\leq \frac{\sqrt{k}\,\|\mathbf{y}-\mathbf{a}_{opt}\|_{2}}{\sigma_{min}(\mathbf{D}_{opt})\,\|\mathbf{a}_{opt}-\mathbf{a}_{j-1}\|_{2}} \\
&\leq \frac{\sqrt{k}\,\|\mathbf{y}-\mathbf{a}_{opt}\|_{2}}{\sqrt{1-d\,\mu_{B1}(k-1)}\,\|\mathbf{a}_{opt}-\mathbf{a}_{j-1}\|_{2}}, \qquad (4.6)
\end{aligned}$$

where $\sigma_{min}(\mathbf{D}_{opt})$ is the minimum singular value of $\mathbf{D}_{opt}$ and the last step is due to Lemma B.4. Note that $\sigma_{min}(\mathbf{D}_{opt})$ is nonzero, since $\mathbf{D}_{opt}$ has full column rank. This is the case due to the assumption $d\,\mu_{B1}(k) < 1/2$. By the properties of the Babel function we have $d\,\mu_{B1}(2k) < 1$ and thus by Proposition 2.5 that the vectors within any k-blocks of $\mathbf{D}$ are linearly independent–thus the columns in $\mathbf{D}_{opt}$ are linearly independent.

We now derive an upper bound on the term $z_{opt}(\mathbf{r}_{j-1})$ in (4.5). Observe that $(\mathbf{a}_{opt}-\mathbf{a}_{j-1}) \in \text{span}(\mathbf{D}_{opt})$ and $\mathbf{D}_{opt}(\mathbf{D}_{opt}^{\dagger})$ is an orthogonal projector on the range of the columns of $\mathbf{D}_{opt}$. Because $\mathbf{D}_{opt}(\mathbf{D}_{opt}^{\dagger})$ is Hermitian, we get

$$\begin{aligned}
z_{opt}(\mathbf{r}_{j-1}) &= \frac{\|\overline{\mathbf{D}}_{opt}^{H}(\mathbf{a}_{opt} - \mathbf{a}_{j-1})\|_{2,\infty}}{\|\mathbf{D}_{opt}^{H}(\mathbf{a}_{opt} - \mathbf{a}_{j-1})\|_{2,\infty}} \\
&= \frac{\|\overline{\mathbf{D}}_{opt}^{H}(\mathbf{D}_{opt}^{\dagger})^{H}\mathbf{D}_{opt}^{H}(\mathbf{a}_{opt} - \mathbf{a}_{j-1})\|_{2,\infty}}{\|\mathbf{D}_{opt}^{H}(\mathbf{a}_{opt} - \mathbf{a}_{j-1})\|_{2,\infty}} \\
&\leq \frac{\|\overline{\mathbf{D}}_{opt}^{H}(\mathbf{D}_{opt}^{\dagger})^{H}\|_{2,\infty}\|\mathbf{D}_{opt}^{H}(\mathbf{a}_{opt} - \mathbf{a}_{j-1})\|_{2,\infty}}{\|\mathbf{D}_{opt}^{H}(\mathbf{a}_{opt} - \mathbf{a}_{j-1})\|_{2,\infty}} \\
&= \|\overline{\mathbf{D}}_{opt}^{H}(\mathbf{D}_{opt}^{\dagger})^{H}\|_{2,\infty}.
\end{aligned}$$

Now, we expand the pseudo inverse as $\mathbf{D}_{opt}^{\dagger} = (\mathbf{D}_{opt}^{H}\mathbf{D}_{opt})^{-1}\mathbf{D}_{opt}^{H}$, which is valid since $\mathbf{D}_{opt}$ has full column rank because of the arguments from above. We get

$$\begin{aligned}
z_{opt}(\mathbf{r}_{j-1}) &\leq \|\overline{\mathbf{D}}_{opt}^{H}[(\mathbf{D}_{opt}^{H}\mathbf{D}_{opt})^{-1}\mathbf{D}_{opt}^{H}]^{H}\|_{2,\infty} \\
&= \|\overline{\mathbf{D}}_{opt}^{H}\mathbf{D}_{opt}(\mathbf{D}_{opt}^{H}\mathbf{D}_{opt})^{-1}\|_{2,\infty} \\
&\leq \underbrace{\|\overline{\mathbf{D}}_{opt}^{H}\mathbf{D}_{opt}\|_{2,\infty}}_{=:\, z_{opt_1}(\mathbf{r}_{j-1})}\underbrace{\|(\mathbf{D}_{opt}^{H}\mathbf{D}_{opt})^{-1}\|_{2,\infty}}_{=:\, z_{opt_2}(\mathbf{r}_{j-1})}. \qquad (4.7)
\end{aligned}$$

We can bound $z_{opt_1}(\mathbf{r}_{j-1})$ above by

$$\begin{aligned}
z_{opt_1}(\mathbf{r}_{j-1}) &= \|\overline{\mathbf{D}}_{opt}^{H}\mathbf{D}_{opt}\|_{2,\infty} \\
&\leq \rho_r\left(\overline{\mathbf{D}}_{opt}^{H}\mathbf{D}_{opt}\right) \\
&= \max_{l}\sum_{r}\rho\left(\overline{\mathbf{D}}_{opt}[l]^{H}\,\mathbf{D}_{opt}[r]\right) \\
&\leq d\,\mu_{B1}(k), \qquad (4.8)
\end{aligned}$$

where the first inequality is due to Lemma 1.1. The last step follows from the definition of the block Babel function.

Now, consider the term $z_{opt_2}(\mathbf{r}_{j-1})$. Define the partial Gram matrix $\mathbf{G}_{opt} := \mathbf{D}_{opt}^{H}\mathbf{D}_{opt}$. Due to orthonormality of the blocks in $\mathbf{D}_{opt}$, it has the form

$$\begin{aligned}
\mathbf{G}_{opt} &= \begin{bmatrix}
\mathbf{I}_d & \mathbf{G}_{opt}[1,2] & \cdots & \cdots & \mathbf{G}_{opt}[1,k] \\
\mathbf{G}_{opt}[2,1] & \mathbf{I}_d & \mathbf{G}_{opt}[2,3] & \cdots & \mathbf{G}_{opt}[2,k] \\
\vdots & \cdots & \ddots & \cdots & \vdots \\
\vdots & \cdots & \cdots & \ddots & \vdots \\
\mathbf{G}_{opt}[k,1] & \cdots & \cdots & \mathbf{G}_{opt}[k,k-1] & \mathbf{I}_d
\end{bmatrix} \\
&=: \mathbf{I} + \mathbf{A},
\end{aligned}$$

with $\mathbf{G}_{opt}[l,r] = (\mathbf{D}_{opt}[l])^{H}\mathbf{D}_{opt}[r]$.

Whenever $\|\mathbf{A}\| < 1$, where $\|\cdot\|$ is some matrix norm, the von Neumann series $\sum_{i=0}^{\infty}(-\mathbf{A})^i$ converges to the inverse of $(\mathbf{I}+\mathbf{A})$ [25]. In our case, consider the mixed matrix norm $\|\mathbf{A}\|_{2,\infty}$ which is bounded above by $d\,\mu_{B1}(k-1) \leq d\,\mu_{B1}(k) < 1/2$. Hence, the von Neumann series converges and we get

$$\begin{aligned} z_{opt_2}(\mathbf{r}_{j-1}) &= \|(\mathbf{D}_{opt}^H \mathbf{D}_{opt})^{-1}\|_{2,\infty} \\ &= \|\sum_{i=0}^{\infty}(-\mathbf{A})^i\|_{2,\infty} \\ &\leq \sum_{i=0}^{\infty}\|\mathbf{A}^i\|_{2,\infty} \\ &\leq \frac{1}{1-\|\mathbf{A}\|_{2,\infty}} \\ &\leq \frac{1}{1-\rho_r(\mathbf{A})} \\ &\leq \frac{1}{1-d\,\mu_{B1}(k-1)} \\ &\leq \frac{1}{1-d\,\mu_{B1}(k)}. \end{aligned} \tag{4.9}$$

Putting (4.8) and (4.9) together yields

$$\begin{aligned} z_{opt}(\mathbf{r}_{j-1}) &\leq z_{opt_1}(\mathbf{r}_{j-1}) \cdot z_{opt_2}(\mathbf{r}_{j-1}) \\ &\leq d\,\mu_{B1}(k)\,\frac{1}{1-d\,\mu_{B1}(k)} \\ &= \frac{d\,\mu_{B1}(k)}{1-d\,\mu_{B1}(k)}. \end{aligned} \tag{4.10}$$

Thus by combining (4.6) and (4.10) we obtain

$$\begin{aligned} z(\mathbf{r}_{j-1}) &\leq z_{err}(\mathbf{r}_{j-1}) + z_{opt}(\mathbf{r}_{j-1}) \\ &\leq \frac{\sqrt{k}\,\|\mathbf{y}-\mathbf{a}_{opt}\|_2}{\sqrt{1-d\,\mu_{B1}(k-1)}\,\|\mathbf{a}_{opt}-\mathbf{a}_{j-1}\|_2} + \frac{d\,\mu_{B1}(k)}{1-d\,\mu_{B1}(k)} \\ &\leq \frac{\sqrt{k}\,\|\mathbf{y}-\mathbf{a}_{opt}\|_2}{\sqrt{1-d\,\mu_{B1}(k)}\,\|\mathbf{a}_{opt}-\mathbf{a}_{j-1}\|_2} + \frac{d\,\mu_{B1}(k)}{1-d\,\mu_{B1}(k)} < 1. \end{aligned} \tag{4.11}$$

We manipulate this to get

$$\frac{\sqrt{k\,[1-d\,\mu_{B1}(k)]}}{1-2\,d\,\mu_{B1}(k)}\|\mathbf{y}-\mathbf{a}_{opt}\|_2 < \|\mathbf{a}_{opt}-\mathbf{a}_{j-1}\|_2.$$

Note that the vectors $(\mathbf{y}-\mathbf{a}_{opt}) \in \ker(\mathbf{D}_{opt})$ and $(\mathbf{a}_{opt}-\mathbf{a}_{j-1}) \in \operatorname{span}(\mathbf{D}_{opt})$ are orthogonal by definition. If they were not orthogonal, $\mathbf{a}_{opt}$ would not be an optimal approximation. We can thus apply Pythagoras' Theorem, that is

$$\begin{aligned}\|\mathbf{y}-\mathbf{a}_{j-1}\|_2 &= \sqrt{\|\mathbf{y}-\mathbf{a}_{opt}\|_2^2+\|\mathbf{a}_{opt}-\mathbf{a}_{j-1}\|_2^2}\\ &> \sqrt{\|\mathbf{y}-\mathbf{a}_{opt}\|_2^2+\left(\frac{\sqrt{k\,[1-d\,\mu_{B1}(k)]}}{1-2\,d\,\mu_{B1}(k)}\right)^2\|\mathbf{y}-\mathbf{a}_{opt}\|_2^2}\\ &= \sqrt{1+\frac{k\,[1-d\,\mu_{B1}(k)]}{[1-2\,d\,\mu_{B1}(k)]^2}}\,\|\mathbf{y}-\mathbf{a}_{opt}\|_2, \qquad (4.12)\end{aligned}$$

which is exactly the statement of the lemma. □

Proof of Proposition 4.1. From Lemma 4.1 we know that BOMP will select another block from an optimal k-block approximation, whenever $d\,\mu_{B1}(k) < 1/2$ and (4.3) holds. Thus, in this case we get into the vicinity of an optimal k-block approximation. That is, the obtained approximation by BOMP after k steps will satisfy

$$\|\mathbf{y}-\mathbf{a}_k\|_2 \leq \sqrt{1+\frac{k\,[1-d\,\mu_{B1}(k)]}{[1-2\,d\,\mu_{B1}(k)]^2}}\,\|\mathbf{y}-\mathbf{a}_{opt}\|_2.$$

□

4.1.2 Theoretical Discussion

BOMP as an Approximation Algorithm

The main result of the previous section is that BOMP can be used as an approximation algorithm to the block-sparse approximation problem.

Let $\mathbf{D}$ be a dictionary with block-size d and block Babel function $\mu_{B1}(\cdot)$. We want to approximate the signal $\mathbf{y} = \mathbf{D}\mathbf{x}$ by a block k-sparse vector $\mathbf{x}_k$ in $\mathbf{D}$. Then, by Corollary 4.1, BOMP is an approximation algorithm to the block-sparse approximation problem for block-sparsity k if $d\,\mu_{B1}(k) < 1/2$ (compare this to the results for OMP in [19]). Assume this condition to be in force. Then, BOMP will always recover a block k-sparse vector $\mathbf{x}_k$ such that the approximation error of the corresponding approximation $\mathbf{a}(\mathbf{x}_k)$ for $\mathbf{y}$ is at most larger by a constant factor than the approximation error of an optimal k-block approximation $\mathbf{a}_{opt}(\mathbf{x}_{opt})$ for $\mathbf{y}$. The actual value of this factor depends on the block-sparsity k and the block Babel function of the dictionary.

As BOMP approximates $\mathbf{y} = \mathbf{D}\mathbf{x}$ by $\mathbf{a}(\mathbf{x}_k)$ if the condition in Lemma 4.1 is satisfied it also approximates $\mathbf{x}$ by $\mathbf{x}_k$.

To quantify the quality of this approximation note that for two vectors $\mathbf{u}, \mathbf{v} \in$

$\mathbb{C}^N$ we have

$$\begin{aligned}\|\mathbf{Du}-\mathbf{Dv}\|_2^2 &= \|\mathbf{D}(\mathbf{u}-\mathbf{v})\|_2^2\\ &= (\mathbf{D}(\mathbf{u}-\mathbf{v}))^H\,(\mathbf{D}(\mathbf{u}-\mathbf{v}))\\ &= (\mathbf{u}-\mathbf{v})^H\mathbf{D}^H\mathbf{D}(\mathbf{u}-\mathbf{v}), \end{aligned} \tag{4.13}$$

where the matrix $\mathbf{D}^H\mathbf{D}$ is positive semidefinite [17]. From (4.2) we get with the above equation that

$$\sqrt{(\mathbf{x}-\mathbf{x}_k)^H\mathbf{D}^H\mathbf{D}(\mathbf{x}-\mathbf{x}_k)} \le c\sqrt{(\mathbf{x}-\mathbf{x}_{opt})^H\mathbf{D}^H\mathbf{D}(\mathbf{x}-\mathbf{x}_{opt})}, \tag{4.14}$$

where

$$c = \sqrt{1+\frac{k\,[1-d\,\mu_{B1}(k)]}{[1-2\,d\,\mu_{B1}(k)]^2}}. \tag{4.15}$$

Thus the vector $\mathbf{x}_k$ recovered by BOMP will always be such that

$$\sqrt{(\mathbf{x}-\mathbf{x}_k)^H\mathbf{D}^H\mathbf{D}(\mathbf{x}-\mathbf{x}_k)}$$

is at most larger by a constant factor c than

$$\sqrt{(\mathbf{x}-\mathbf{x}_{opt})^H\mathbf{D}^H\mathbf{D}(\mathbf{x}-\mathbf{x}_{opt})}.$$

Note however that the mapping $f:\mathbb{C}^N\to\mathbb{R}$ such that

$$f(\mathbf{z}) \mapsto \sqrt{\mathbf{z}^H\mathbf{D}^H\mathbf{D}\mathbf{z}}$$

in general defines a semi-norm and a not a norm.

Comparison to the Sparse Case

In this section we make some comments on signal approximation with OMP and BOMP and on the relation of Proposition 4.1 to the corresponding result for OMP (see Section 2.1.5 for details, originally from [19]).

For this assume that we want to approximate some signal $\mathbf{y}$ by a k-block approximation $\mathbf{a}(\mathbf{x}_{\text{BOMP}})$ or a kd-term approximation $\mathbf{a}'(\mathbf{x}_{\text{OMP}})$ in the dictionary $\mathbf{D}$ with block-size d. The dictionary has Babel function $\mu_1(\cdot)$ and block Babel function $\mu_{B1}(\cdot)$.

We obtain the block k-sparse vector $\mathbf{x}_{\text{BOMP}}$ by BOMP in k-steps and the kd-sparse vector $\mathbf{x}_{\text{OMP}}$ by OMP in kd-steps. It is then interesting to compare the approximation errors of the approximations $\mathbf{a}(\mathbf{x}_{\text{BOMP}})$ and $\mathbf{a}'(\mathbf{x}_{\text{OMP}})$ for $\mathbf{y}$. For this we make use of Proposition 4.1 and 2.3. That is, we compare the equations

$$\|\mathbf{y}-\mathbf{a}(\mathbf{x}_{\text{BOMP}})\|_2 \le c\,\|\mathbf{y}-\mathbf{a}_{opt}\|_2, \tag{4.16}$$

where

$$c = \sqrt{1 + \frac{k\,[1 - d\,\mu_{B1}(k)]}{[1 - 2\,d\,\mu_{B1}(k)]^2}}, \tag{4.17}$$

and

$$\|\mathbf{y} - \mathbf{a}(\mathbf{x}_{\mathrm{OMP}})\|_2 \leq c'\,\|\mathbf{y} - \mathbf{a}'_{opt}\|_2, \tag{4.18}$$

where

$$c' = \sqrt{1 + \frac{kd\,[1 - \mu_1(kd)]}{[1 - 2\,\mu_1(kd)]^2}}. \tag{4.19}$$

An optimal k-block approximation of $\mathbf{y}$ is $\mathbf{a}_{opt}$, while $\mathbf{a}'_{opt}$ is an optimal kd-term approximation of $\mathbf{y}$. We say c and c' are the *approximation error constants* for a given dictionary $\mathbf{D}$, block-size d and block-sparsity k.

There is no obvious way to relate (4.16) and (4.18). However, we can make some comments:

- OMP does not approximate a solution to the block-sparse approximation problem but it approximates a solution to the sparse approximation problem.

- Whenever there is no optimal kd-term approximation that exhibits a block structure in the dictionary $\mathbf{D}$, then the minimal approximation error $\|\mathbf{y} - \mathbf{a}'_{opt}\|_2$ in the sparse case is smaller than the minimal approximation error $\|\mathbf{y} - \mathbf{a}_{opt}\|_2$ in the block-sparse case, i.e., $\|\mathbf{y} - \mathbf{a}'_{opt}\|_2 \leq \|\mathbf{y} - \mathbf{a}_{opt}\|_2$.

  Proof. Assume $\mathbf{a}_{opt}$ is an optimal k-block approximation for $\mathbf{y}$ and $\mathbf{a}'_{opt}$ is an optimal kd-term approximation for $\mathbf{y}$. Further, assume that $\|\mathbf{y} - \mathbf{a}'_{opt}\|_2 > \|\mathbf{y} - \mathbf{a}_{opt}\|_2$. But then $\mathbf{a}'_{opt}$ cannot be an optimal kd-term approximation since $\mathbf{a}_{opt}$ is also a kd-term approximation but has a smaller approximation error. Hence, there is a contradiction. □

- For a fixed number of blocks k, and the corresponding number of terms kd to recover, the approximation error constant c can be smaller than c' in some cases. Assume now that $c < c'$. Then we are guaranteed to get within the smaller factor, namely the factor c, to an optimal k-block approximation when using BOMP to recover k blocks. On the other hand, when using OMP to recover kd coefficients we are guaranteed to get within the factor c' to an optimal kd-term approximation.

  Comparing the approximation error constants c and c' resorts mainly to the comparison of the Babel function and the block Babel function, respectively. Unfortunately, there seems to be no obvious relation between the block Babel function and the Babel function of a dictionary.

Another fact to take into account when comparing the approximation error constants for OMP and BOMP in the previous described environment is the following: Despite the difference of the block Babel function and the Babel function, the approximation error constants depend on k and kd, respectively. But $k < kd$ in our scenario, whenever $d > 1$.

4.1.3 Simulation Results

The aim of this section is to demonstrate that BOMP can be used as an approximation algorithm to the block-sparse approximation problem. We show that signal approximation by BOMP can be advantageous over signal approximation by OMP when the signal to recover exhibits block-structure and is observed by noisy measurements.

Signal Recovery from Noisy Measurements

Consider the experiment of recovering an unknown block k-sparse vector $\mathbf{x}$ from noisy measurements $\mathbf{y}$ in the dictionary $\mathbf{D}$ with block-size d. That is,

$$\mathbf{y} = \mathbf{D}\mathbf{x} + \mathbf{z},$$

where $\mathbf{z}$ is a noise vector. With BOMP we recover a block k-sparse representation $\mathbf{x}_{\text{BOMP}}$ in k-steps and with OMP we recover a kd-sparse representation $\mathbf{x}_{\text{OMP}}$ in kd-steps from $\mathbf{y}$ and $\mathbf{D}$. To relate the quality of the obtained representations we compare $\|\mathbf{x}-\mathbf{x}_{\text{BOMP}}\|_2$ and $\|\mathbf{x}-\mathbf{x}_{\text{OMP}}\|_2$, i.e., the mismatch energy of the obtained representations and the real representation. For a fair comparison over multiple experiments we normalized the mismatch energy by the energy of $\mathbf{x}$, i.e., by $\|\mathbf{x}\|_2$.

We carried out this experiment for fixed and random dictionaries of varying dimensions and with a varying number of contained blocks. In all cases we chose the block-sparsity k of $\mathbf{x}$ as the maximum possible value such that the condition in Proposition 4.1 is satisfied. Then we created $\mathbf{x}$ by selecting k blocks uniformly at random and drawing the coefficients within these blocks from an i.i.d. Gaussian distribution. All other coefficients of the vector are set to zero.

The coefficients of the noise vector $\mathbf{z}$ are drawn from an i.i.d. Gaussian distribution with zero mean. The variance of this noise vector is chosen such that the Signal to Noise Ratio (SNR), given as

$$\text{SNR} := \frac{\mathbb{E}(\|\mathbf{D}\mathbf{x}\|_2)}{\mathbb{E}(\|\mathbf{z}\|_2)}, \tag{4.20}$$

where $\mathbb{E}(\cdot)$ is the expectation operator, takes some specific value.

In the experiments performed with random dictionaries, we created the dictionaries as described in Section 3.2.3 with a slight modification: The random dictionary $\mathbf{D}_i$ is created from the random matrix $\mathbf{A}_i$ by setting the lth block of

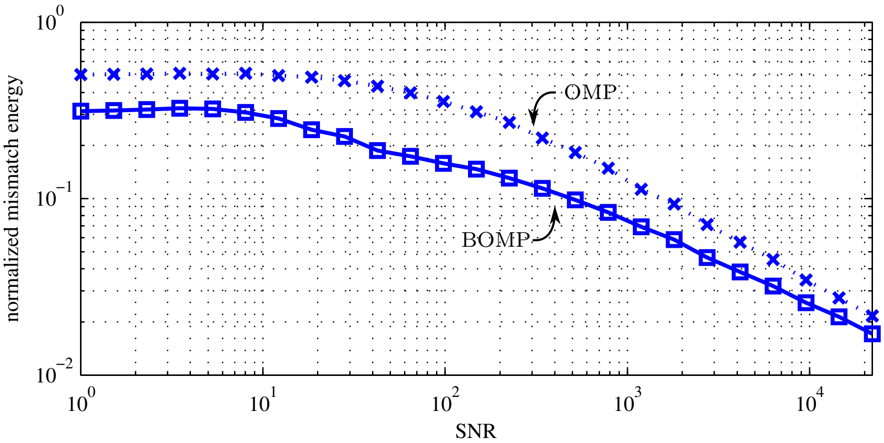


Figure 4.1: Recovery performance of OMP and BOMP for signals embedded in noise against the SNR. The plots show the normalized mismatch energy for the fixed dictionary $\mathbf{D} = [\mathbf{I}_M, \mathbf{F}_{M/d} \otimes \mathbf{I}_d]$, where $M = 1024$ and the block-size $d = 8$.

$\mathbf{D}_i$ to the orthonormal block created from the lth block of $\mathbf{A}_i$ by the MATLAB command `orth`, for $l = 1, \ldots, R$.

For each considered value of the SNR we averaged the mismatch energy of the obtained representations over 100 pairs of realizations of the random dictionary (or fixed dictionary) and the unknown random vector.

The results of these experiments are shown in Figures 4.1 and 4.2. Observe that BOMP obtains representations that are closer, in the ℓ_2-norm, to the original representation than the representations obtained by OMP.

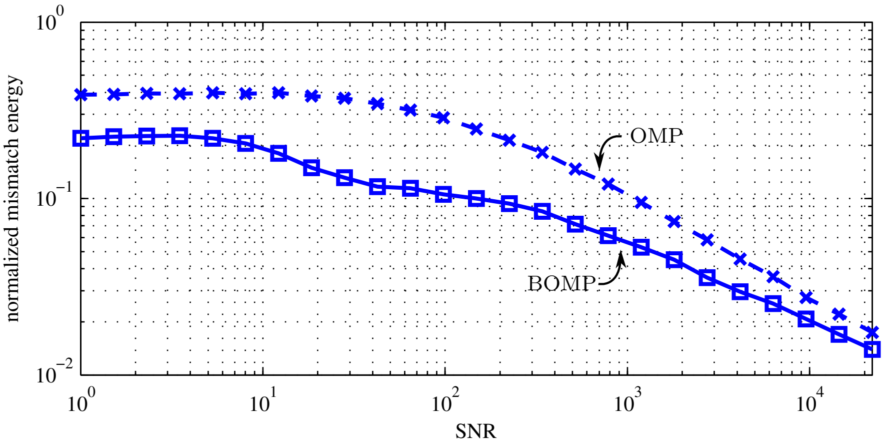


Figure 4.2: Recovery performance of OMP and BOMP for signals embedded in noise against the SNR. The plots show the normalized mismatch energy for random dictionaries of size $M \times N$, where $M = 1024, N = 2048$ and the block-size $d = 8$.

Chapter 5

Conclusions

In this thesis we gave an overview of compressed sensing for the regular sparse case and of compressed sensing under the block-sparsity model assumption. We covered topics like dictionary analysis, unique sparse representation and recovery methods.

As our contributions we generalized some results from the regular sparse case to the block-sparse case and showed that we can achieve better results in this way—in detail:

- **Unique Block-Sparse Representation.** We derived a sufficient condition for unique block-sparse representation of signals in dictionaries comprised of orthonormal bases. This condition can guarantee unique block-sparse representation for higher block-sparsities than similar conditions for arbitrary dictionaries. Furthermore, it is advantageous over similar conditions for the regular sparse case if the block-sparsity model is applicable.

- **Signal Recovery.** In Section 3.2 we derived a sufficient condition for exact signal recovery with L-OPT and BOMP in dictionaries comprised of orthonormal bases. This condition can guarantee the recovery of signals with a higher block-sparsity than ensured by similar conditions for arbitrary dictionaries.

- **Block-Sparse Approximation.** We justified the usage of BOMP as an approximation algorithm to the block-sparse approximation problem. Furthermore, we showed by simulations that signal recovery with BOMP can be advantageous over signal recovery with OMP if the nonzero coefficients of the signal to be recovered occur in clusters (blocks) and the signal is observed through noisy measurements.

Bibliography

[1] H. Nyquist, “Certain topics in telegraph transmission theory,” *Proceedings of the IEEE*, vol. 90, no. 2, pp. 280–305, 2002.

[2] C. Shannon, “Communication in the presence of noise,” *Proceedings of the IEEE*, vol. 86, no. 2, pp. 447–457, February 1998.

[3] J. Romberg, “Imaging via compressive sampling [introduction to compressive sampling and recovery via convex programming],” *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 14–20, March 2008.

[4] D. S. Taubman and M. W. Marcellin, *Jpeg2000: Image Compression Fundamentals, Standards, and Practice.* Kluwer Academic Pulishers, August 2001.

[5] R. Baraniuk, “Compressive sensing,” *Lecture Notes in IEEE Signal Processing Magazine,*, vol. 24, pp. 118–120, July 2007.

[6] D. Donoho, “Compressed sensing,” *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, April 2006.

[7] D. L. Donoho, M. Vetterli, R. A. Devore, and I. Daubechies, “Data compression and harmonic analysis,” *IEEE Transactions on Information Theory*, vol. 44, pp. 2435–2476, 1998.

[8] J. Mairal, G. Sapiro, and M. Elad, “Multiscale sparse image representationwith learned dictionaries,” in *Image Processing, 2007. ICIP 2007. IEEE International Conference on*, vol. 3, October 2007, pp. III–105–III–108.

[9] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde, “Model-based compressive sensing,” *IEEE Transactions on Information Theory*, 2008, submitted.

[10] Y. C. Eldar and M. Mishali, “Robust recovery of signals from a structured union of subspaces,” *IEEE Transactions on Information Theory*, 2008, submitted.

[11] M. Mishali and Y. Eldar, “Reduce and boost: Recovering arbitrary sets of jointly sparse vectors,” *IEEE Transactions on Signal Processing*, vol. 56, no. 10, pp. 4692–4702, October 2008.

[12] ——, “Blind multiband signal reconstruction: Compressed sensing for analog signals,” *IEEE Transactions on Signal Processing*, vol. 57, no. 3, pp. 993–1009, March 2009.

[13] M. Elad and A. Bruckstein, “A generalized uncertainty principle and sparse representation in pairs of bases,” *IEEE Transactions on Information Theory*, vol. 48, no. 9, pp. 2558–2567, September 2002.

[14] R. Gribonval and M. Nielsen, “Sparse representations in unions of bases,” *IEEE Transactions on Information Theory*, vol. 49, no. 12, pp. 3320–3325, December 2003.

[15] H. Rauhut, K. Schnass, and P. Vandergheynst, “Compressed sensing and redundant dictionaries,” *IEEE Transactions on Information Theory*, vol. 54, no. 5, pp. 2210–2219, May 2008.

[16] Y. Eldar, P. Kuppinger, and H. Bölcskei, “Compressed sensing of block-sparse signals: Uncertainty relations and efficient recovery,” *IEEE Transactions on Signal Processing*, June 2009, submitted.

[17] R. Horn and C. Johnson, *Matrix Analysis.* Cambridge University Press, February 1990.

[18] E. J. Candes and M. B. Wakin, “An introduction to compressive sampling,” *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 21–30, March 2008.

[19] J. Tropp, “Greed is good: algorithmic results for sparse approximation,” *IEEE Transactions on Information Theory*, vol. 50, no. 10, pp. 2231–2242, October 2004.

[20] S. Chen, D. L. Donoho, and M. A. Saunders, “Atomic decomposition by basis pursuit,” *SIAM Journal on Scientific Computing*, vol. 20, pp. 33–61, 1998.

[21] S. G. Mallat and Z. Zhang, “Matching pursuits with time-frequency dictionaries,” *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, December 1993.

[22] D. Donoho and Y. Tsaig, “Fast solution of ℓ_1 -norm minimization problems when the solution may be sparse,” *IEEE Transactions on Information Theory*, vol. 54, no. 11, pp. 4789–4812, November 2008.

[23] M. S. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret, “Applications of second-order cone programming,” *Linear Algebra and its Applications*, vol. 284, no. 1-3, pp. 193–228, November 1998.

[24] T. Cover and J. Thomas, *Elements of Information Theory.* Wiley-Interscience, August 1991.

[25] E. E. Tyrtyshnikov, *A Brief Introduction to Numerical Analysis.* Birkhï¿œser Boston, July 1997.

[26] J. Laub, *Matrix Analysis for Scientists and Engineers.* Society for Industrial Mathematics, December 2004.

[27] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 5, pp. 2197–2202, March 2003.

[28] S. W. N. Jorge, *Numerical Optimization.* New York, USA: Springer, August 1999.

Appendix A

Symbols Used

For a list of symbols used throughout this thesis see Table A.1.

Symbol	Description
$\mathbb{C}$	The set of complex numbers.
$\mathbb{R}$	The set of real numbers.
$\mathbb{R}_+$	The set of nonnegative real numbers.
$\mathcal{X}_k$	Set of block k-sparse vectors of given length N.
$\mathcal{Y}_k$	Set of signals that have a block k-sparse representation in a given dictionary $\mathbf{D}$.
$\mathrm{supp}(\mathbf{x})$	Support of the vector $\mathbf{x}$.
$\mathrm{supp}_B(\mathbf{x})$	Block-support of the vector $\mathbf{x}$.
$\delta[i]$	Kronecker Delta.
$\lvert a\rvert$	Absolute value of a.
$\lvert S\rvert$	Cardinality of the set S.
$\mathbf{F}_N$	DFT-matrix of size $N \times N$.
$\mathbf{I}_N$	Identity matrix of size $N \times N$.
$\mathbf{0}_N$	All zero matrix of size $N \times N$ or all zero vector of length N.
$\tilde{\mathbf{I}}_{k_1d\times k_2d}, \tilde{\mathbf{I}}_{kd}$	For k, k_1, k_2, d integers with d the block-size, $\tilde{\mathbf{I}}_{k_1d\times k_2d}$ denotes the matrix $$\begin{bmatrix} \mathbf{I}_d & \dots & \mathbf{I}_d \\ \vdots & \ddots & \vdots \\ \mathbf{I}_d & \dots & \mathbf{I}_d \end{bmatrix}$$ of size $k_1d \times k_2d$. The matrix $\tilde{\mathbf{I}}_{kd} = \tilde{\mathbf{I}}_{kd\times kd}$.
$\tilde{\mathbf{1}}_{kd}$	For k, d integers with d the block-size, $\tilde{\mathbf{1}}_{kd} = \tilde{\mathbf{I}}_{kd\times d}$.
$\mathrm{span}(\mathbf{A})$	Column space of the matrix $\mathbf{A}$.
$\mathbf{A} \otimes \mathbf{B}$	Kronecker product of the matrices $\mathbf{A}$ and $\mathbf{B}$.

Continued on next page

Symbol	Description
$\rho(\mathbf{A})$	Spectral norm of the square matrix $\mathbf{A}$.
$\langle \mathbf{x}, \mathbf{y} \rangle$	Inner product of $\mathbf{x}$ and $\mathbf{y}$.
$\|\mathbf{x}\|_0$	Number of nonzero coefficients of the vector $\mathbf{x}$; sparsity of the vector $\mathbf{x}$.
$\|\mathbf{x}\|_p$	ℓ_p-norm of the vector $\mathbf{x}$.
$\|\mathbf{x}\|_{2,0}$	Block-sparsity of the vector $\mathbf{x}$, i.e., the number of nonzero blocks in $\mathbf{x}$.
$\|\mathbf{x}\|_{2,p}$	Mixed ℓ_2/ℓ_p-norm of the vector $\mathbf{x}$.
$\|\mathbf{A}\|_p$	Induced ℓ_p-norm of the matrix $\mathbf{A}$.
$\|\mathbf{A}\|_{2,p}$	Mixed ℓ_2/ℓ_p-norm of the matrix $\mathbf{A}$.
$\|\mathbf{A}\|_{\max}$	Elementwise norm with $p = \infty$, i.e., $\|\mathbf{A}\|_{\max} = \max \lvert A_{i,j} \rvert$.
$\mathbf{x}^H$	Conjugate transpose of the vector $\mathbf{x}$.
$\mathbf{A}^H$	Conjugate transpose of the matrix $\mathbf{A}$.
$\mathbf{A}^\dagger$	Pseudo-inverse of the matrix $\mathbf{A}$.
$Z_B(\mathbf{A})$	Block-spark of the matrix $\mathbf{A}$.
$\ker(\mathbf{A})$	Kernel of the matrix $\mathbf{A}$.
$\mu(\mathbf{D}), \mu$	Coherence of the dictionary $\mathbf{D}$.
$\mu_1(\mathbf{D}, m), \mu_1(m)$	Babel function of the dictionary $\mathbf{D}$.
$\mu_B(\mathbf{D}), \mu_B$	Block-Coherence of the dictionary $\mathbf{D}$.
$\mu_{B1}(\mathbf{D}, m), \mu_{B1}(m)$	Block Babel function of the dictionary $\mathbf{D}$.

Table A.1: List of used symbols

Appendix B

Mathematical Background

B.1 Mixed Norm Properties

Lemma B.1. *The mixed norms for matrices, defined as*

$$\|\mathbf{A}\|_{2,p} := \max_{\mathbf{x}\neq 0} \frac{\|\mathbf{A}\mathbf{x}\|_{2,p}}{\|\mathbf{x}\|_{2,p}},$$

are submultiplicative. That is, for any two matrices $\mathbf{A}$ *and* $\mathbf{B}$ *(of appropriate dimensions for matrix multiplication) we have*

$$\|\mathbf{AB}\|_{2,p} \leq \|\mathbf{A}\|_{2,p}\|\mathbf{B}\|_{2,p}.$$

Proof. If $\mathbf{AB} = \mathbf{0}$ the statement trivially follows. Hence, consider the case $\mathbf{AB} \neq \mathbf{0}$. Then, submultiplicity follows from the following calculation:

$$\begin{aligned}
\|\mathbf{AB}\|_{2,p} &= \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{ABx}\|_{2,p}}{\|\mathbf{x}\|_{2,p}} = \max_{\mathbf{x}\neq\mathbf{0},\mathbf{Bx}\neq\mathbf{0}} \frac{\|\mathbf{A}(\mathbf{Bx})\|_{2,p}\,\|\mathbf{Bx}\|_{2,p}}{\|\mathbf{x}\|_{2,p}\,\|\mathbf{Bx}\|_{2,p}} \\
&\leq \max_{\mathbf{Bx}\neq\mathbf{0}} \frac{\|\mathbf{ABx}\|_{2,p}}{\|\mathbf{Bx}\|_{2,p}} \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{Bx}\|_{2,p}}{\|\mathbf{x}\|_{2,p}} \\
&\leq \max_{\mathbf{y}\neq\mathbf{0}} \frac{\|\mathbf{Ay}\|_{2,p}}{\|\mathbf{y}\|_{2,p}} \max_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{Bx}\|_{2,p}}{\|\mathbf{x}\|_{2,p}} \\
&= \|\mathbf{A}\|_{2,p}\,\|\mathbf{B}\|_{2,p}\,,
\end{aligned}$$

where we used the substitution $\mathbf{y} := \mathbf{Bx}$ in the second last step. □

B.1.1 Inequalities

Lemma B.2. *Let* $\mathbf{A} \in \mathbb{C}^{M\times N}$ *be a matrix, where* $M = md$, m, d *and* N *are integers. This matrix can be partitioned as*

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 \\ \vdots \\ \mathbf{A}_m \end{bmatrix},$$

where $\mathbf{A}_i \in \mathbb{C}^{d\times N}$, *i.e.,* $\mathbf{A}_i$ *is the ith d-row block of matrix* $\mathbf{A}$. *Then*

$$\|\mathbf{A}\|_{2,\infty} \geq \|\mathbf{A}_i\|_2 \quad \textit{for } i = 1,\ldots,m. \tag{B.1}$$

Proof. By the definition of the mixed matrix norm we have

$$\begin{aligned}\|\mathbf{A}\|_{2,\infty} &= \max\{\|\mathbf{Ax}\|_{2,\infty} : \|\mathbf{x}\|_{2,\infty} \leq 1\} \\ &= \max\{\max_j \|\mathbf{A}_j\mathbf{x}\|_2 : \|\mathbf{x}\|_{2,\infty} \leq 1\} \\ &\geq \max\{\|\mathbf{A}_i\mathbf{x}\|_2 : \|\mathbf{x}\|_{2,\infty} \leq 1\} \text{ for } i = 1,\ldots,m, \\ &\geq \max\{\|\mathbf{A}_i\mathbf{x}\|_2 : \|\mathbf{x}\|_2 \leq 1\} \text{ for } i = 1,\ldots,m, \\ &= \|\mathbf{A}_i\|_2 \text{ for } i = 1,\ldots,m.\end{aligned}$$

The last inequality follows from the fact that the constraint $\|\mathbf{x}\|_{2,\infty} \leq 1$ is less restrictive than the constraint $\|\mathbf{x}\|_2 \leq 1$, i.e.,

$$\{\mathbf{Ax} : \|\mathbf{x}\|_2 \leq 1\} \subseteq \{\mathbf{Ax} : \|\mathbf{x}\|_{2,\infty} \leq 1\}.$$

□

Lemma B.3. *Let* $\mathbf{A} \in \mathbb{C}^{M\times N}$ *be a matrix, where* $M = md$, m, d, *and* N *are integers. Then*

$$\|\mathbf{A}\|_{2,\infty} \geq \frac{1}{\sqrt{m}}\|\mathbf{A}\|_2. \tag{B.2}$$

Proof. The matrix $\mathbf{A}$ can be partitioned as

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 \\ \vdots \\ \mathbf{A}_m \end{bmatrix},$$

with $\mathbf{A}_i \in \mathbb{C}^{d\times N}$ for $i = 1,\ldots,m$, i.e., $\mathbf{A}_i$ is the ith d-row block of $\mathbf{A}$.

Then, there is some index j such that

$$\alpha := \|\mathbf{A}_j\|_2 \geq \|\mathbf{A}_i\|_2, \text{ for } i = 1,\ldots,m.$$

Due to the definition of the induced matrix norm $\|\cdot\|_2$, there is some $\mathbf{x}$ with $\|\mathbf{x}\|_2 = 1$ such that $\|\mathbf{A}_j\mathbf{x}\|_2 = \alpha$. We thus have

$$\|\mathbf{A}\|_2 = \max_{\|\mathbf{x}'\|_2=1} \left\| \begin{matrix} \mathbf{A}_1\mathbf{x}' \\ \vdots \\ \mathbf{A}_m\mathbf{x}' \end{matrix} \right\|_2 \leq \left\| \begin{matrix} \alpha \\ \vdots \\ \alpha \end{matrix} \right\|_2 = \sqrt{m}\,\alpha = \sqrt{m}\,\|\mathbf{A}_j\|_2. \tag{B.3}$$

On the other hand, consider the mixed ℓ_2/ℓ_∞-norm of $\mathbf{A}$, that, by Lemma B.2, bounds the ℓ_2-norm of any d-row block $\mathbf{A}_i$, i.e.,

$$\|\mathbf{A}\|_{2,\infty} \geq \alpha = \|\mathbf{A}_j\|_2. \tag{B.4}$$

Thus by combining inequality (B.3) and (B.4) we have

$$\|\mathbf{A}\|_{2,\infty} \geq \alpha \geq \frac{1}{\sqrt{m}}\|\mathbf{A}\|_2 \tag{B.5}$$

and the statement follows. □

B.2 Gershgorin Disc Theorem

In this section we prove a block version of the Gershgorin disc theorem (compare to [17] for the classical version of theorem).

Lemma B.4 (Adopted from [19], Lemma 2.3)**.** *Let* $\mathbf{D}$ *be a matrix consisting of* m *orthonormal blocks of size* d*. Then, the singular values of* $\mathbf{D}$ *exceed* $[1 - d\,\mu_{B1}(m-1)]$*, where* $\mu_{B1}(\cdot)$ *is the block Babel function of* $\mathbf{D}$*.*

Proof. This proof closelfy follows the proof of the Gershgorin Theorem in [17]. Consider the generalized gram matrix $\mathbf{G} = \mathbf{D}^H\mathbf{D}$. Then the eigenvalues of $\mathbf{G}$ are the squared singular values of $\mathbf{D}$.

Now, let λ be any of the eigenvalues of $\mathbf{G}$ and $\mathbf{x} \neq 0$ the corresponding normalized eigenvector. We then have $\mathbf{G}\mathbf{x} = \lambda\mathbf{x}$ and for each block $(\mathbf{G}\mathbf{x})[j] = \lambda\mathbf{x}[j]$.

We choose i such that $\|\mathbf{x}[i]\|_2 \geq \|\mathbf{x}[j]\|_2$ for any possible index j, i.e., i is the index of the block of $\mathbf{x}$ with the largest ℓ_2-norm. Then we rewrite $(\mathbf{G}\mathbf{x})[i] = \lambda\mathbf{x}[i]$ as

$$\sum_j \mathbf{G}[i,j]\,\mathbf{x}[j] = \lambda\mathbf{x}[i]\,,$$

where $\mathbf{G}[i,j] = \mathbf{D}[i]^H\,\mathbf{D}[j]$.

Then we subtract $\mathbf{G}[i,i]\,\mathbf{x}[i]$ to get

$$\begin{aligned} \sum_{j\neq i} \mathbf{G}[i,j]\,\mathbf{x}[j] &= \lambda\mathbf{x}[i] - \mathbf{G}[i,i]\,\mathbf{x}[i] \\ &= (\lambda\mathbf{I} - \mathbf{G}[i,i])\mathbf{x}[i]\,. \end{aligned}$$

Now take the ℓ_2-norm of both sides and divide by $\|\mathbf{x}[i]\|_2$. This yields

$$\frac{\|\sum_{j\neq i}\mathbf{G}[i,j]\,\mathbf{x}[j]\|_2}{\|\mathbf{x}[i]\|_2} = \frac{\|(\lambda\mathbf{I}-\mathbf{G}[i,i])\mathbf{x}[i]\|_2}{\|\mathbf{x}[i]\|_2}.$$

Applying the triangle inequality and the submultiplicity of the ℓ_2-norm to the left side of the above equation yields

$$\begin{aligned}\frac{\|(\lambda\mathbf{I}-\mathbf{G}[i,i])\mathbf{x}[i]\|_2}{\|\mathbf{x}[i]\|_2} &\leq \frac{\sum_{j\neq i}\|\mathbf{G}[i,j]\,\mathbf{x}[j]\|_2}{\|\mathbf{x}[i]\|_2}\\ &\leq \frac{\sum_{j\neq i}\|\mathbf{G}[i,j]\|_2\|\mathbf{x}[j]\|_2}{\|\mathbf{x}[i]\|_2}\\ &\leq \sum_{j\neq i}\|\mathbf{G}[i,j]\|_2\\ &\leq d\,\mu_{B1}(m-1),\end{aligned}$$

where the second last inequality is because of the selection of the index i.

Now consider the left hand side of the above inequality. Since the blocks of $\mathbf{D}$ are orthonormal by assumption (see Section 2.1), $\mathbf{G}[i,i]$ is an identity matrix of size $d\times d$. Thus the expression $\|(\lambda\mathbf{I}-\mathbf{G}[i,i])\mathbf{x}[i]\|_2$ can be simplified as $|\lambda-1|\|\mathbf{x}[i]\|_2$. Hence the above equation can be rewritten as

$$|\lambda-1| = |1-\lambda| \leq d\,\mu_{B1}(m-1)$$

and the statement follows. □

Appendix C

Block-Sparsity in Arbitrary Dictionaries

C.1 Block-Coherence Inequality

Lemma C.1. *Let* $\mathbf{D}$ *be a dictionary comprised of orthogonal blocks of size* d *with coherence* μ *and block-coherence* μ_B*. The block-coherence then satisfies*

$$\frac{\mu}{d} \leq \mu_B \leq \mu. \tag{C.1}$$

Proof. The upper bound is due to [16].

To derive the lower bound, consider the generalized gram matrix $\mathbf{G} = \mathbf{D}^H \mathbf{D}$. There is a $d \times d$ submatrix $\mathbf{G}[l,r] = \mathbf{D}[l]^H \, \mathbf{D}[r]$, with $l \neq r$, with an entry of absolute value μ because of the orthogonal blocks in the dictionary and the definition of the coherence.

By the matrix norm inequality

$$\|\mathbf{A}\|_{max} \leq \rho(\mathbf{A}),$$

where $\mathbf{A}$ is a square matrix and $\|\mathbf{A}\|_{\max}$ is the maximum absolute value of any matrix element of $\mathbf{A}$ [26], we thus have

$$\|\mathbf{G}[l,r]\|_{max} = \mu \leq \rho(\mathbf{G}[l,r]) \leq \mu_B \, d.$$

Hence, $\mu/d \leq \mu_B$. □

C.2 Proof of Proposition 2.5

Before stating the proof of Proposition 2.5 we present the following helpful lemma.

Lemma C.2 (Similar to Theorem 4 in [27]). *Let* $\mathbf{D}$ *be a dictionary consisting of orthonormal blocks of size* d *and with block Babel function* $\mu_{B1}(m)$. *A necessary condition for some nonzero vector* $\mathbf{z}$ *with* $\|\mathbf{z}\|_{2,0} = m'$ *to be in the kernel of* $\mathbf{D}$ *is*

$$m' \geq \min\{m : d\,\mu_{B1}(m-1) \geq 1\}. \tag{C.2}$$

Consequently, the block-spark of $\mathbf{D}$ *obeys*

$$Z_B(\mathbf{D}) \geq \min\{m : d\,\mu_{B1}(m-1) \geq 1\}.$$

Proof. Assume $\mathbf{z} \in \ker(\mathbf{D})\backslash\{\mathbf{0}\}$ and $\|\mathbf{z}\|_{2,0} = m'$. Let $\mathcal{I}$ be the block-support of $\mathbf{z}$, i.e., $\mathcal{I} = \operatorname{supp}_B(\mathbf{z})$. Further, let i_0 be any element of $\mathcal{I}$ and let $\mathcal{J} = \mathcal{I}\backslash\{i_0\}$. Then $|\mathcal{I}| = m'$ and $|\mathcal{J}| = m' - 1$.

Since $\mathbf{z} \in \ker(\mathbf{D})$ we have $\mathbf{0} = \mathbf{D}\mathbf{z}$. Thus, we can rewrite this as

$$\mathbf{z}[i_0] = -\sum_{i\in\mathcal{J}} \mathbf{D}[i_0]^H\, \mathbf{D}[i]\, \mathbf{z}[i].$$

Applying the ℓ_2-norm, the triangle inequality and the submultiplicity yields

$$\|\mathbf{z}[i_0]\|_2 \leq \sum_{i\in\mathcal{J}} \|\mathbf{D}[i_0]^H\, \mathbf{D}[i]\|_2 \|\mathbf{z}[i]\|_2.$$

We add $\|\mathbf{D}[i_0]^H\, \mathbf{D}[i_0]\|_2\|\mathbf{z}[i_0]\|_2 = \|\mathbf{z}[i_0]\|_2$ (as the blocks are orthonormal) to both sides to get

$$2\|\mathbf{z}[i_0]\|_2 \leq \sum_{i\in\mathcal{I}} \|\mathbf{D}[i_0]^H\, \mathbf{D}[i]\|_2 \|\mathbf{z}[i]\|_2.$$

Now, we sum over all possible $i_0 \in \mathcal{I}$, that is

$$2\|\mathbf{z}\|_{2,1} \leq \sum_{i_0\in\mathcal{I}} \sum_{i\in\mathcal{I}} \|\mathbf{D}[i_0]^H\, \mathbf{D}[i]\|_2 \|\mathbf{z}[i]\|_2.$$

From this, we obtain

$$\begin{aligned}2\|\mathbf{z}\|_{2,1} &\leq \sum_{i_0\in\mathcal{I}} \sum_{i\in\mathcal{I}} \|\mathbf{D}[i_0]^H\, \mathbf{D}[i]\|_2 \|\mathbf{z}[i]\|_2\\ &= \sum_{i\in\mathcal{I}} \|\mathbf{z}[i]\|_2 \sum_{i_0\in\mathcal{I}} \|\mathbf{D}[i_0]^H\, \mathbf{D}[i]\|_2.\end{aligned}$$

Bounding the right hand side of the above equation by the block Babel function yields

$$\begin{aligned}2\|\mathbf{z}\|_{2,1} &\leq \sum_{i\in\mathcal{I}} \|\mathbf{z}[i]\|_2 \sum_{i_0\in\mathcal{I}} \|\mathbf{D}[i_0]^H\, \mathbf{D}[i]\|_2\\ &\leq \sum_{i\in\mathcal{I}} \|\mathbf{z}[i]\|_2 [d\,\mu_{B1}(m'-1)+1]\\ &= [d\,\mu_{B1}(m'-1)+1]\|\mathbf{z}\|_{2,1},\end{aligned}$$

where the term $\mathbf{D}[i_0]^H \mathbf{D}[i_0]$ is treated separately.

We simplify this to

$$1 \leq d\,\mu_{B1}(m' - 1).$$

Hence, any nonzero vector in the kernel of the dictionary must have at least $\min\{m : d\,\mu_{B1}(m-1) \geq 1\}$ nonzero blocks. That is, the block-spark of $\mathbf{D}$ obeys

$$Z_B(\mathbf{D}) \geq \min\{m : d\,\mu_{B1}(m-1) \geq 1\}.$$

□

Proof of Proposition 2.5. Use Lemma C.2 and apply Lemma 3.2 to obtain the statement. □

C.3 Proof of Proposition 2.6

This proof closely follows the path taken in [14].

Before stating the proof, we need to extend parts of the program laid out in Section 3.1. Therefore, we state a lemma similar to Lemma 3.1.

Lemma C.3 (Adopted from [14], Lemma 1)**.** *Let* $\mathbf{D}$ *be a dictionary and* $S \subset \Delta_B = \{1, \ldots, R\}$ *be a set of block-indices. Define*

$$P_{B,1}(S, \mathbf{D}) := \max_{\substack{\mathbf{z} \in \ker(\mathbf{D}) \\ \mathbf{z} \neq \mathbf{0}}} \frac{\sum\limits_{l \in S} \|\mathbf{z}[l]\|_2}{\sum\limits_{l} \|\mathbf{z}[l]\|_2}. \tag{C.3}$$

If $P_{B,1}(S, \mathbf{D}) < 1/2$*, then, for all* $\mathbf{x}$ *such that* $\mathbf{y} = \mathbf{D}\mathbf{x}$ *and* $\mathit{supp}_B(\mathbf{x}) \subset S$*,* $\mathbf{x}$ *is the unique solution to the problem* (2.34)*.*

Proof. The proof is analogous to the proof of Lemma 3.1. □

The next lemma will be helpful in proving Proposition 2.6.

Lemma C.4 (Adopted from [14], Lemma 2)**.** *If* $|S| < k$ *is a sufficient condition for* $P_{B,1}(S, \mathbf{D}) < 1/2$*, then it is also a sufficient condition for* $P_{B,0}(S, \mathbf{D}) < 1/2$*.*

Proof. Refer to the proof of Lemma 2 in [14]. □

Now we have all prerequisites to prove Proposition 2.6.

Proof of Proposition 2.6. This proof closely follows the proof of Theorem 1 in [14].

We first show that $|S| < 1/2\,[1{+}1/(d\,\mu_B)]$ is a sufficient condition for $P_{B,1}(S, \mathbf{D}) < 1/2$ and then conclude the statement by applying Lemma C.3 and Lemma C.4.

For any $\mathbf{z} \in \ker(\mathbf{D})$ we have $\mathbf{0} = \mathbf{D}\mathbf{z}$. Hence, for some fixed l

$$\mathbf{D}[l]\,\mathbf{z}[l] = -\sum_{k \neq l} \mathbf{D}[k]\,\mathbf{z}[k].$$

We multiply this from the left with $\mathbf{D}[l]^H$. Since the blocks of $\mathbf{D}$ are orthonormal by assumption, we get

$$\mathbf{z}[l] = -\sum_{k \neq l} \mathbf{D}[l]^H\,\mathbf{D}[k]\,\mathbf{z}[k].$$

Taking the ℓ_2-norm of both sides and applying the triangle inequality yields

$$\|\mathbf{z}[l]\|_2 \leq \sum_{k \neq l} \|\mathbf{D}[l]^H\,\mathbf{D}[k]\,\mathbf{z}[k]\|_2.$$

By using the submultiplicity of the ℓ_2-norm and the fact that the spectral norm of $\mathbf{D}[l]^H\,\mathbf{D}[k]$ is bounded above by $d\,\mu_B$ we obtain

$$\|\mathbf{z}[l]\|_2 \leq d\,\mu_B \sum_{k \neq l} \|\mathbf{z}[k]\|_2.$$

It follows that

$$(1 + d\,\mu_B)\,\|\mathbf{z}[l]\|_2 \leq d\,\mu_B \sum_{k} \|\mathbf{z}[k]\|_2.$$

Now we sum over all $l \in S$, where S is any set of block-indices, to get

$$(1 + d\,\mu_B) \sum_{l \in S} \|\mathbf{z}[l]\|_2 \leq |S|\,d\,\mu_B \sum_{k} \|\mathbf{z}[k]\|_2.$$

Hence,

$$P_{B,1}(S, \mathbf{D}) \leq |S|\,\frac{d\,\mu_B}{(1 + d\,\mu_B)}.$$

We need the right hand side of above equation to be less than $1/2$ to ensure that $P_{B,1}(S, \mathbf{D}) < 1/2$. This is guaranteed if

$$|S| < \frac{1}{2}\left(1 + \frac{1}{d\,\mu_B}\right).$$

Hence, if we have some $\mathbf{y} = \mathbf{D}\mathbf{x}$ with $\|\mathbf{x}\|_{2,0} < 1/2\,[1 + 1/(d\,\mu_B)]$, then $\mathbf{x}$ uniquely solves the minimization problem (2.34).

By Lemma C.4 and Lemma C.3 we know that $\|\mathbf{x}\|_{2,0} < 1/2\,[1 + 1/(d\,\mu_B)]$ is also sufficient for $\mathbf{x}$ to uniquely solve the problem (2.27). □

Appendix D

Proof of Proposition 3.2

The proof of Proposition 3.2 (or actually the proof of Lemma D.2 which we need as an intermediate step) is long and cumbersome. Before stating the proof we will derive two helpful lemmas.

Further, there is Lemma D.3 which we will need to guarantee the convergence of some von Neumann series in our derivation. To not distract the reader too much from our goal of proving Proposition 3.2 the lemma and its proof are postponed to the end of this section.

At some points in the derivation we will require a matrix $\mathbf{A}$ of size $Ed \times Fd$, where E, F, d are integers and d is the block-size, to have the following form:

$$\mathbf{A} = \begin{bmatrix} \alpha_{1,1}^{\mathbf{A}} \mathbf{I}_d & \dots & \alpha_{1,F}^{\mathbf{A}} \mathbf{I}_d \\ \vdots & \ddots & \vdots \\ \alpha_{E,1}^{\mathbf{A}} \mathbf{I}_d & \dots & \alpha_{E,F}^{\mathbf{A}} \mathbf{I}_d \end{bmatrix}, \tag{D.1}$$

with $\alpha_{l,r}^{\mathbf{A}} \in \mathbb{R}$.

We will also make use of the following Lemma:

Lemma D.1. *Let $\mathbf{A} \in \mathbb{C}^{Ed \times Kd}$ and $\mathbf{B} \in \mathbb{C}^{Kd \times Fd}$ be arbitrary matrices, where E, F, K, d are integers and d is the block-size. Then*

$$\rho((\mathbf{AB})[l,r]) \leq \rho\big((\mathbf{A}'\mathbf{B}')[l,r]\big), \tag{D.2}$$

for all $1 \leq l \leq E$ and $1 \leq r \leq F$, where $\mathbf{A}', \mathbf{B}'$ are of form (D.1)*, $\alpha_{l,r}^{\mathbf{A}'} \geq \rho(\mathbf{A}[l,r])$ and $\alpha_{l,r}^{\mathbf{B}'} \geq \rho(\mathbf{B}[l,r])$. Consequently,*

$$\rho_c(\mathbf{AB}) \leq \rho_c\big(\mathbf{A}'\mathbf{B}'\big). \tag{D.3}$$

Proof. The (l,r)-th $d \times d$ submatrix of the matrix $\mathbf{AB}$ is given by

$$(\mathbf{AB})[l,r] = \sum_{k=1}^{K} \mathbf{A}[l,k]\,\mathbf{B}[k,r].$$

We can upper bound the spectral norm of each of this blocks by

$$\begin{aligned}\rho((\mathbf{AB})[l,r]) &= \rho\left(\sum_{k=1}^{K} \mathbf{A}[l,k]\,\mathbf{B}[k,r]\right)\\ &\leq \sum_{k=1}^{K} \rho(\mathbf{A}[l,k])\,\rho(\mathbf{B}[k,r])\,,\end{aligned}$$

where we used the triangle inequality and submultiplicity of the spectral norm.

But

$$\sum_{k=1}^{K} \rho(\mathbf{A}[l,k])\,\rho(\mathbf{B}[k,r]) \leq \sum_{k=1}^{K} \rho\big(\mathbf{A}'[l,k]\big)\,\rho\big(\mathbf{B}'[k,r]\big)\,,$$

because of the construction of $\mathbf{A}'$ and $\mathbf{B}'$.

To conclude the proof observe that

$$\begin{aligned}\sum_{k=1}^{K} \rho\big(\mathbf{A}'[l,k]\big)\,\rho\big(\mathbf{B}'[k,r]\big) &= \sum_{k=1}^{K} \rho\Big(\alpha_{l,k}^{\mathbf{A}'}\mathbf{I}_d\Big)\,\rho\Big(\alpha_{k,r}^{\mathbf{B}'}\mathbf{I}_d\Big)\\ &= \rho\left(\sum_{k=1}^{K} \alpha_{l,k}^{\mathbf{A}'}\mathbf{I}_d \alpha_{k,r}^{\mathbf{B}'}\mathbf{I}_d\right)\\ &= \rho\left(\sum_{k=1}^{K} \mathbf{A}'[l,k]\,\mathbf{B}'[k,r]\right)\\ &= \rho\big((\mathbf{A}'\mathbf{B}')[l,r]\big)\,.\end{aligned}$$

□

The following lemma generalizes a result from [19] (originally from [14]) to the block-sparse case.

Lemma D.2 (Adopted from [19], Theorem 3.7)**.** *Let the dictionary* $\mathbf{D}$ *consist of* $L \geq 2$ *orthonormal bases and let* d *be the block-size. The dictionary has block-coherence* μ_B*. Assume the signal* $\mathbf{y}$ *to have a unique block-sparsest representation* $\mathbf{x}$ *with block-sparsity* $k = n_1 + \cdots + n_L$*. The numbers* n_l *denote the number of blocks from basis* l *corresponding to the nonzero blocks in* $\mathbf{x}$*. Assume them to be ordered, that is* $0 \leq n_1 \leq \ldots \leq n_L$*. If*

$$\sum_{l=2}^{L} \frac{d\,\mu_B\,n_l}{1 + d\,\mu_B\,n_l} < \frac{1}{2\,(1 + d\,\mu_B\,n_1)}, \tag{D.4}$$

then the Exact Recovery Condition (2.37) *is obeyed and* $\mathbf{x}$ *can be recovered by L-OPT and BOMP.*

Proof of Lemma D.2. This proof closely follows the path taken in [19].

By assumption $\mathbf{y} = \mathbf{D}\mathbf{x}$, where $\mathbf{x}$ is block k-sparse. Let $\mathbf{D}_{opt}$ consist of the blocks of $\mathbf{D}$ corresponding to the nonzero blocks of $\mathbf{x}$, and let $\overline{\mathbf{D}}_{opt}$ consist of the blocks of $\mathbf{D}$ that are not in $\mathbf{D}_{opt}$. Further, let the blocks within $\mathbf{D}_{opt}$ be arranged such that $\mathbf{D}_{opt} = [\mathbf{D}_{opt_1} \ \ldots \ \mathbf{D}_{opt_L}]$, where the n_l blocks of the matrix $\mathbf{D}_{opt_l}$ are from the lth basis. We develop an upper bound on the quantity

$$\rho_c\Big(\mathbf{D}_{opt}^{\dagger}\overline{\mathbf{D}}_{opt}\Big)$$

occurring in the Exact Recovery Condition and then conclude the statement from it.

We start by expanding the pseudo inverse of $\mathbf{D}_{opt}$ in the above expression, that is

$$\begin{aligned}\rho_c\Big(\mathbf{D}_{opt}^{\dagger}\overline{\mathbf{D}}_{opt}\Big) &= \rho_c\big([\mathbf{D}_{opt}^{H}\mathbf{D}_{opt}]^{-1}\mathbf{D}_{opt}^{H}\overline{\mathbf{D}}_{opt}\big)\\ &= \max_{l} \rho_c\big([\mathbf{D}_{opt}^{H}\mathbf{D}_{opt}]^{-1}\mathbf{D}_{opt}^{H}\overline{\mathbf{D}}_{opt}[l]\big) \qquad \text{(D.5)}\end{aligned}$$

This expansion is valid since $\mathbf{D}_{opt}$ has full column rank—this is the case because of $\mathbf{x}$ being the unique sparsest representation of $\mathbf{y}$.

Now we construct a matrix $\mathbf{E}$ from $\mathbf{G}_1 := \mathbf{D}_{opt}^{H}\mathbf{D}_{opt}$ such that $\mathbf{E}^{-1}$ is of form (D.1) and that

$$\rho\big(\mathbf{G}_1^{-1}[l,r]\big) \leq \rho\big(\mathbf{E}^{-1}[l,r]\big) \qquad \text{(D.6)}$$

for all l, r. Further, we construct a matrix $\mathbf{F}$ from $\mathbf{G}_2 := \mathbf{D}_{opt}^{H}\overline{\mathbf{D}}_{opt}[l]$ such that $\mathbf{F}$ is of form (D.1) and such that

$$\rho(\mathbf{G}_2[l,r]) \leq \rho(\mathbf{F}[l,r]) \qquad \text{(D.7)}$$

for all l, r. Then, by Lemma D.1,

$$\rho_c\big(\mathbf{G}_1^{-1}\mathbf{G}_2\big) \leq \rho_c\big(\mathbf{E}^{-1}\mathbf{F}\big)\,.$$

Construction of E. The term $\mathbf{G}_1 = \mathbf{D}_{opt}^{H}\mathbf{D}_{opt}$ can be written as

$$\mathbf{G}_1 = \begin{bmatrix} \mathbf{I}_{n_1 d} & -\mathbf{A}_{1,2} & \ldots & -\mathbf{A}_{1,L} \\ -\mathbf{A}_{2,1} & \mathbf{I}_{n_2 d} & \ldots & -\mathbf{A}_{2,L} \\ \vdots & \vdots & \ddots & \vdots \\ -\mathbf{A}_{L,1} & \ldots & -\mathbf{A}_{L,L-1} & \mathbf{I}_{n_L d} \end{bmatrix} =: \mathbf{I}_{kd} - \mathbf{A},$$

with $\mathbf{A}_{i,j} = -\mathbf{D}_{opt_i}^{H}\mathbf{D}_{opt_j} \in \mathbb{C}^{n_i d \times n_j d}$. In more detail, we have

$$\mathbf{A}_{i,j} = \begin{bmatrix} \mathbf{A}_{i,j}[1,1] & \ldots & \mathbf{A}_{i,j}[1,n_j] \\ \vdots & \ddots & \vdots \\ \mathbf{A}_{i,j}[n_i,1] & \ldots & \mathbf{A}_{i,j}[n_i,n_j] \end{bmatrix},$$

where $\mathbf{A}_{i,j}[k,l] = -\mathbf{D}_{opt}[(n_1 + \ldots + n_{i-1}) + k]^H \, \mathbf{D}_{opt}[(n_1 + \ldots + n_{j-1}) + l]$.

Every $d \times d$ submatrix $\mathbf{A}_{i,j}[k,l]$ of $\mathbf{A}$ has a spectral norm that is less than or equal to $d\,\mu_B$ due to the definition of the block-coherence.

Now we define $\mathbf{E} := \mathbf{I}_{kd} - \mathbf{A}'$, where

$$\mathbf{A}' = \begin{bmatrix} \mathbf{0}_{n_1 d} & \mathbf{A}'_{1,2} & \ldots & \mathbf{A}'_{1,L} \\ \mathbf{A}'_{2,1} & \mathbf{0}_{n_2 d} & \ldots & \mathbf{A}'_{2,L} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{A}'_{L,1} & \ldots & \mathbf{A}'_{L,L-1} & \mathbf{0}_{n_L d} \end{bmatrix}$$

with

$$\mathbf{A}'_{i,j} = d\,\mu_B\,\tilde{\mathbf{I}}_{n_i d \times n_j d}$$

and

$$\tilde{\mathbf{I}}_{n_i d \times n_j d} = \begin{bmatrix} \mathbf{I}_d & \ldots & \mathbf{I}_d \\ \vdots & \ddots & \vdots \\ \mathbf{I}_d & \ldots & \mathbf{I}_d \end{bmatrix} \in \mathbb{C}^{n_i d \times n_j d}.$$

To see that by the above definition inequality (D.6) is obeyed and that $\mathbf{E}^{-1}$ is of form (D.1) we rewrite the inverses of $\mathbf{G}_1$ and $\mathbf{E}$ as von Neumann series [17]. This is valid since $\rho_c(\cdot)$ is a matrix norm, $\rho_c(\mathbf{A}) \leq \rho_c(\mathbf{A}') \leq (k - n_1)\,d\,\mu_B \leq k\,d\,\mu_B$ and by Lemma D.3. Hence,

$$\begin{aligned} \mathbf{G}_1^{-1} &= \mathbf{I}_{kd} + \sum_{i=1}^{\infty} \mathbf{A}^i, \text{ and} \\ \mathbf{E}^{-1} &= \mathbf{I}_{kd} + \sum_{i=1}^{\infty} \mathbf{A}'^i. \end{aligned}$$

By this

$$\begin{aligned} \rho\big(\mathbf{G}_1^{-1}[l,r]\big) &= \rho\left(\delta[l-r]\,\mathbf{I}_d + \left(\sum_{i=1}^{\infty} \mathbf{A}^i\right)[l,r]\right) \\ &\leq \rho(\delta[l-r]\,\mathbf{I}_d) + \rho\left(\left(\sum_{i=1}^{\infty} \mathbf{A}^i\right)[l,r]\right) \\ &\leq \rho(\delta[l-r]\,\mathbf{I}_d) + \rho\left(\left(\sum_{i=1}^{\infty} \mathbf{A}'^i\right)[l,r]\right) \\ &= \rho\big((\mathbf{E}^{-1})[l,r]\big), \end{aligned}$$

where the inequalities follow from the triangle inequality and the submultiplicity of the spectral norm.

We now evaluate $\mathbf{E}^{-1}$. For this, note that we can rewrite $\mathbf{E} = \mathbf{I}_{kd} - \mathbf{A}'$ as

$$\mathbf{E} = (\mathbf{I}_{kd} + d\,\mu_B\,\mathbf{B}) - (\mathbf{A}' + d\,\mu_B\,\mathbf{B}),$$

where $\mathbf{B}$ is constructed as

$$\mathbf{B} = \begin{bmatrix} \tilde{\mathbf{I}}_{n_1 d} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \tilde{\mathbf{I}}_{n_2 d} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & \mathbf{0} & \tilde{\mathbf{I}}_{n_L d} \end{bmatrix} \in \mathbb{C}^{kd \times kd},$$

with the definition $\tilde{\mathbf{I}}_{n_i d} := \tilde{\mathbf{I}}_{n_i d \times n_i d}$.

Note that $\mathbf{A}' + d\,\mu_B\,\mathbf{B} = d\,\mu_B\,\tilde{\mathbf{I}}_{kd}$. Hence,

$$\begin{aligned} \mathbf{E}^{-1} &= \Big((\mathbf{I}_{kd} + d\,\mu_B\,\mathbf{B}) - d\,\mu_B\,\tilde{\mathbf{I}}_{kd}\Big)^{-1} \\ &= \Big(\mathbf{I}_{kd} - d\,\mu_B\,(\mathbf{I}_{kd} + d\,\mu_B\,\mathbf{B})^{-1}\tilde{\mathbf{I}}_{kd}\Big)^{-1}(\mathbf{I}_{kd} + d\,\mu_B\,\mathbf{B})^{-1}, \end{aligned} \tag{D.8}$$

where the last step is because $(\mathbf{C}\mathbf{D})^{-1} = \mathbf{D}^{-1}\mathbf{C}^{-1}$ for invertible matrices $\mathbf{C}$ and $\mathbf{D}$.

We continue by working out the inverse of $(\mathbf{I}_{kd} + d\,\mu_B\,\mathbf{B})^{-1}$ with the von Neumann series—this is valid since $\rho_c(d\,\mu_B\,\mathbf{B}) = d\,\mu_B\,n_L \leq d\,\mu_B\,k < 1$ which guarantees convergence of the series. In this way,

$$(\mathbf{I}_{kd} + d\,\mu_B\,\mathbf{B})^{-1} = \begin{bmatrix} \mathbf{T}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{T}_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & \mathbf{0} & \mathbf{T}_L \end{bmatrix},$$

where

$$\mathbf{T}_i := \mathbf{I}_{n_i d} - \frac{d\,\mu_B}{1 + d\,\mu_B\,n_i}\,\tilde{\mathbf{I}}_{n_i d} \text{ for } i = 1, \dots, L,$$

because

$$\begin{aligned}
\mathbf{T}_i &= \mathbf{I}_{n_i d} + \sum_{i=1}^{\infty} (-d\,\mu_B\,\tilde{\mathbf{I}}_{n_i d})^i \\
&= \mathbf{I}_{n_i d} + \sum_{i=1}^{\infty} (-d\,\mu_B)^i\, n_i^{i-1}\,\tilde{\mathbf{I}}_{n_i d} \\
&= \mathbf{I}_{n_i d} + \tilde{\mathbf{I}}_{n_i d}\,\frac{1}{n_i}\sum_{i=1}^{\infty} (-d\,\mu_B\,n_i)^i \\
&= \mathbf{I}_{n_i d} + \tilde{\mathbf{I}}_{n_i d}\,\frac{1}{n_i}\sum_{i=0}^{\infty} (-d\,\mu_B\,n_i)^i - \frac{1}{n_i}\,\tilde{\mathbf{I}}_{n_i d} \\
&= \mathbf{I}_{n_i d} + \tilde{\mathbf{I}}_{n_i d}\,\frac{1}{n_i}\left(\frac{1}{1+d\,\mu_B\,n_i} - \frac{1+d\,\mu_B\,n_i}{1+d\,\mu_B\,n_i}\right) \\
&= \mathbf{I}_{n_i d} - \frac{d\,\mu_B}{1+d\,\mu_B\,n_i}\,\tilde{\mathbf{I}}_{n_i d}.
\end{aligned}$$

We now investigate the first term in (D.8). Note that we can again rewrite this as a von Neumann series, i.e.,

$$\begin{aligned}
&\Big(\mathbf{I}_{kd} - d\,\mu_B\,(\mathbf{I}_{kd} + d\,\mu_B\,\mathbf{B})^{-1}\,\tilde{\mathbf{I}}_{kd}\Big)^{-1} = \\
&\mathbf{I}_{kd} + \sum_{i=1}^{\infty}\Big(d\,\mu_B\,(\mathbf{I}_{kd} + d\,\mu_B\,\mathbf{B})^{-1}\,\tilde{\mathbf{I}}_{kd}\Big)^i. \qquad \text{(D.9)}
\end{aligned}$$

This is valid since

$$\begin{aligned}
\rho_c\Big(d\,\mu_B\,(\mathbf{I}_{kd} + d\,\mu_B\,\mathbf{B})^{-1}\,\tilde{\mathbf{I}}_{kd}\Big) &= d\,\mu_B\left[n_L\left(1 - \frac{d\,\mu_B\,n_L}{1+d\,\mu_B\,n_L}\right)\right] \\
&= d\,\mu_B\,n_L - \frac{d^2\,\mu_B^2\,n_L^2}{1+d\,\mu_B\,n_L} \\
&< 1.
\end{aligned}$$

Now we compute the term $d\,\mu_B\,(\mathbf{I}_{kd} + d\,\mu_B\,\mathbf{B})^{-1}\,\tilde{\mathbf{I}}_{kd}$ in (D.9) with the von

Neumann series as

$$\begin{aligned} d\,\mu_B\,(\mathbf{I}_{kd} + d\,\mu_B\,\mathbf{B})^{-1}\,\tilde{\mathbf{I}}_{kd} &= d\,\mu_B \begin{bmatrix} \tilde{\mathbf{I}}_{n_1 d\times kd} - \frac{d\,\mu_B\,n_1}{1+d\,\mu_B\,n_1}\,\tilde{\mathbf{I}}_{n_1 d\times kd} \\ \vdots \\ \tilde{\mathbf{I}}_{n_L d\times kd} - \frac{d\,\mu_B\,n_L}{1+d\,\mu_B\,n_L}\,\tilde{\mathbf{I}}_{n_L d\times kd} \end{bmatrix} \\ &= \begin{bmatrix} \frac{d\,\mu_B}{1+d\,\mu_B\,n_1}\,\tilde{\mathbf{I}}_{n_1 d\times kd} \\ \vdots \\ \frac{d\,\mu_B}{1+d\,\mu_B\,n_L}\,\tilde{\mathbf{I}}_{n_L d\times kd} \end{bmatrix} \\ &= \begin{bmatrix} \frac{d\,\mu_B}{1+d\,\mu_B\,n_1}\,\tilde{\mathbf{1}}_{n_1 d} \\ \vdots \\ \frac{d\,\mu_B}{1+d\,\mu_B\,n_L}\,\tilde{\mathbf{1}}_{n_L d} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{1}}^T_{n_1 d} & \dots & \tilde{\mathbf{1}}^T_{n_L d} \end{bmatrix} \\ &= \mathbf{v}\,\tilde{\mathbf{1}}^T_{kd}, \end{aligned}$$

with

$$\tilde{\mathbf{1}}^T_{jd} = \begin{bmatrix} \mathbf{I}_d & \dots & \mathbf{I}_d \end{bmatrix} \in \mathbb{C}^{d\times jd}.$$

Hence, (D.9) becomes

$$\begin{aligned} &\mathbf{I}_{kd} + \sum_{i=1}^{\infty} \left(d\,\mu_B\,(\mathbf{I}_{kd} + d\,\mu_B\,\mathbf{B})^{-1}\,\tilde{\mathbf{I}}_{kd} \right)^i \\ &= \mathbf{I}_{kd} + \sum_{i=1}^{\infty} \left(\mathbf{v}\,\tilde{\mathbf{1}}^T_{kd}\right)^i \\ &= \mathbf{I}_{kd} + \sum_{i=1}^{\infty} \underbrace{\mathbf{v}\,\tilde{\mathbf{1}}^T_{kd}\cdots\mathbf{v}\,\tilde{\mathbf{1}}^T_{kd}}_{i\text{ times}} \\ &= \mathbf{I}_{kd} + \sum_{i=1}^{\infty} \mathbf{v}\left(\tilde{\mathbf{1}}^T_{kd}\,\mathbf{v}\right)^{i-1}\tilde{\mathbf{1}}^T_{kd} \\ &= \mathbf{I}_{kd} + \sum_{i=1}^{\infty} \left(\sum_{l=1}^{L} \frac{d\,\mu_B n_l}{1+d\,\mu_B\,n_l} \right)^{i-1} \mathbf{v}\,\mathbf{I}_d\,\tilde{\mathbf{1}}^T_{kd} \\ &= \mathbf{I}_{kd} + \left(\sum_{i=1}^{\infty} \sum_{l=1}^{L} \left(\frac{d\,\mu_B\,n_l}{1+d\,\mu_B\,n_l} \right)^{i-1} \right) \mathbf{v}\,\tilde{\mathbf{1}}^T_{kd} \\ &= \mathbf{I}_{kd} + \frac{1}{1 - \sum_{l=1}^{L} \left(\frac{d\,\mu_B\,n_l}{1+d\,\mu_B\,n_l} \right)} \mathbf{v}\,\tilde{\mathbf{1}}^T_{kd}, \end{aligned}$$

where the fourth step is because

$$\tilde{\mathbf{1}}^T_{kd}\,\mathbf{v} = \sum_{l=1}^{L} \frac{d\,\mu_B\,n_l}{1+d\,\mu_B\,n_l}\,\mathbf{I}_d.$$

We have now evaluated both terms occurring in the inverse of the matrix $\mathbf{E}$. By this

$$\mathbf{E}^{-1} = \left(\mathbf{I}_{kd} + \frac{1}{1 - \sum\limits_{l=1}^{L} \left(\frac{d\,\mu_B\,n_l}{1+d\,\mu_B\,n_l} \right)} \mathbf{v}\,\tilde{\mathbf{1}}_{kd}^T \right) (\mathbf{I}_{kd} + d\,\mu_B\,\mathbf{B})^{-1} \tag{D.10}$$

Construction of F. We construct the matrix $\mathbf{F}$ from the term $\mathbf{G}_2 = \mathbf{D}_{opt}^H \overline{\mathbf{D}}_{opt}[l]$. Assume that the block $\overline{\mathbf{D}}_{opt}[l]$ is from basis z. Then

$$\mathbf{G}_2 = \left[\ (\mathbf{D}_{opt}{}_1^H \overline{\mathbf{D}}_{opt}[l])^T \quad \dots \quad \mathbf{0}_{n_z d \times d}^T \quad \dots \quad (\mathbf{D}_{opt}{}_L^H \overline{\mathbf{D}}_{opt}[l])^T \ \right]^T,$$

because the block $\overline{\mathbf{D}}_{opt}[l]$ is orthogonal to all other blocks of the same basis.

Define

$$\mathbf{F} := \left[\ d\,\mu_B\,\tilde{\mathbf{1}}_{n_1 d}^T \quad \dots \quad \mathbf{0}_{n_z d \times d}^T \quad \dots \quad d\,\mu_B\,\tilde{\mathbf{1}}_{n_L d}^T \ \right]^T. \tag{D.11}$$

With this $\mathbf{F}$ is of form (D.1) and

$$\rho(\mathbf{G}_2[l,r]) \le \rho(\mathbf{F}[l,r])$$

for all l, r.

Putting things together. As mentioned before, for the matrices $\mathbf{E}$ and $\mathbf{F}$ we have

$$\rho_c\big(\mathbf{G}_1^{-1}\mathbf{G}_2\big) \le \rho_c\big(\mathbf{E}^{-1}\mathbf{F}\big).$$

Now we investigate the product $\mathbf{E}^{-1}\mathbf{F}$ in detail.

From the last term of $\mathbf{E}^{-1}$ and $\mathbf{F}$ we get

$$\begin{aligned} &(\mathbf{I}_{kd} + d\,\mu_B\,\mathbf{B})^{-1} \left[\ d\,\mu_B\,\tilde{\mathbf{1}}_{n_1 d}^T \quad \dots \quad \mathbf{0}_{n_z d \times d}^T \quad \dots \quad d\,\mu_B\,\tilde{\mathbf{1}}_{n_L d}^T \ \right]^T \\ &\quad = \left[\ \frac{d\,\mu_B}{1+d\,\mu_B\,n_1}\,\tilde{\mathbf{1}}_{n_1 d}^T \quad \dots \quad \mathbf{0}_{n_z d \times d}^T \quad \dots \quad \frac{d\,\mu_B}{1+d\,\mu_B\,n_L}\,\tilde{\mathbf{1}}_{n_L d}^T \ \right]^T. \end{aligned} \tag{D.12}$$

Hence,

$$\mathbf{E}^{-1}\mathbf{F} = \begin{bmatrix} \frac{d\,\mu_B}{1+d\,\mu_B\,n_1}\,\tilde{\mathbf{1}}_{n_1 d} \\ \vdots \\ \mathbf{0}_{n_z d \times d} \\ \vdots \\ \frac{d\,\mu_B}{1+d\,\mu_B\,n_L}\,\tilde{\mathbf{1}}_{n_L d} \end{bmatrix} + \frac{\sum\limits_{l=1, l\neq z}^{L} \frac{d\,\mu_B\,n_l}{1+d\,\mu_B\,n_l}}{1 - \sum\limits_{l=1}^{L} \frac{d\,\mu_B\,n_l}{1+d\,\mu_B\,n_l}} \begin{bmatrix} \frac{d\,\mu_B}{1+d\,\mu_B\,n_1}\,\tilde{\mathbf{1}}_{n_1 d} \\ \vdots \\ \frac{d\,\mu_B}{1+d\,\mu_B\,n_z}\,\tilde{\mathbf{1}}_{n_z d} \\ \vdots \\ \frac{d\,\mu_B}{1+d\,\mu_B\,n_L}\,\tilde{\mathbf{1}}_{n_L d} \end{bmatrix}.$$

With this, we finally get

$$
\begin{aligned}
\rho_c\Big(\mathbf{D}_{opt}^{\dagger}\overline{\mathbf{D}}_{opt}\Big) &\leq \rho_c\big(\mathbf{E}^{-1}\mathbf{F}\big) \\
&= \sum_{l=1, l\neq z}^{L} \frac{d\,\mu_B\, n_l}{1 + d\,\mu_B\, n_l} \left(1 + \frac{\sum\limits_{l=1, l\neq z}^{L} \frac{d\,\mu_B\, n_l}{1+d\,\mu_B\, n_l}}{1 - \sum\limits_{l=1}^{L} \frac{d\,\mu_B\, n_l}{1+d\,\mu_B\, n_l}} \right) \\
&\quad + \frac{d\,\mu_B\, n_z}{1 + d\,\mu_B\, n_z} \frac{\sum\limits_{l=1, l\neq z}^{L} \frac{d\,\mu_B\, n_l}{1+d\,\mu_B\, n_l}}{1 - \sum\limits_{l=1}^{L} \frac{d\,\mu_B\, n_l}{1+d\,\mu_B\, n_l}} \\
&= \sum_{l=1, l\neq z}^{L} \frac{d\,\mu_B\, n_l}{1 + d\,\mu_B\, n_l} \left(\frac{1}{1 - \sum\limits_{l=1}^{L} \frac{d\,\mu_B\, n_l}{1+d\,\mu_B\, n_l}} - \frac{\frac{d\,\mu_B\, n_z}{1+d\,\mu_B\, n_z}}{1 - \sum\limits_{l=1}^{L} \frac{d\,\mu_B\, n_l}{1+d\,\mu_B\, n_l}} \right) \\
&\quad + \frac{d\,\mu_B\, n_z}{1 + d\,\mu_B\, n_z} \frac{\sum\limits_{l=1, l\neq z}^{L} \frac{d\,\mu_B\, n_l}{1+d\,\mu_B\, n_l}}{1 - \sum\limits_{l=1}^{L} \frac{d\,\mu_B\, n_l}{1+d\,\mu_B\, n_l}} \\
&= \frac{\sum\limits_{l=1, l\neq z}^{L} \frac{d\,\mu_B\, n_l}{1+d\,\mu_B\, n_l}}{1 - \sum\limits_{l=1}^{L} \frac{d\,\mu_B\, n_l}{1+d\,\mu_B\, n_l}}.
\end{aligned}
$$

The bound on $\rho_c\Big(\mathbf{D}_{opt}^{\dagger}\overline{\mathbf{D}}_{opt}\Big)$ is weakest for $z = 1$, since the corresponding basis contributes the fewest blocks to the optimal representation.

We derive the original statement of the proof since the Exact Recovery Condition requires $\rho_c\Big(\mathbf{D}_{opt}^{\dagger}\overline{\mathbf{D}}_{opt}\Big) < 1$ and we just apply this bound to the above inequality. Hence we end up with

$$
\sum_{l=2}^{L} \frac{d\,\mu_B\, n_l}{1 + d\,\mu_B\, n_l} < \frac{1}{2\,(1 + d\,\mu_B\, n_1)} \tag{D.13}
$$

as a sufficient condition for successful signal recovery with L-OPT and BOMP. □

Proof of Proposition 3.2. Follow the arguments in [14]. □

Lemma D.3. *Let* $\mathbf{D}$*,* L*,* $n_1, \ldots, n_L$ *and* k *be as in Lemma D.2. If*

$$
\sum_{l=2}^{L} \frac{d\,\mu_B\, n_l}{1 + d\,\mu_B\, n_l} < \frac{1}{2\,(1 + d\,\mu_B\, n_1)}, \tag{D.14}
$$

then $k\,d\,\mu_B < 1$.

Proof. We split this proof into two parts. First, we consider the case $L \geq 3$ and then the case $L = 2$.

More than two orthonormal bases ($L \geq 3$). In this case we get the statement directly from Condition (D.14). Rewrite the condition as

$$\sum_{l=2}^{L} \frac{d\,\mu_B\,n_l}{1 + d\,\mu_B\,n_l} - \frac{1}{2\,(1 + d\,\mu_B\,n_1)} < 0,$$

and multiply it by $(1 + d\,\mu_B\,n_L)$ to get

$$\sum_{l=2}^{L} \frac{d\,\mu_B\,n_l(1 + d\,\mu_B\,n_L)}{1 + d\,\mu_B\,n_l} - \frac{1 + d\,\mu_B\,n_L}{2\,(1 + d\,\mu_B\,n_1)} < 0.$$

Observe that

$$\sum_{l=2}^{L} \frac{d\,\mu_B\,n_l\,(1 + d\,\mu_B\,n_L)}{1 + d\,\mu_B\,n_l} \geq \sum_{l=2}^{L} d\,\mu_B\,n_l,$$

since $n_L \geq n_l$ for $l = 1, \ldots, L$, to obtain

$$\sum_{l=2}^{L} d\,\mu_B\,n_l - \frac{1 + d\,\mu_B\,n_L}{2\,(1 + d\,\mu_B\,n_1)} < 0.$$

Now, multiply this with $2\,(1 + d\,\mu_B\,n_1)$ to get

$$2\sum_{l=2}^{L} d\,\mu_B\,n_l + 2\sum_{l=2}^{L} (d\,\mu_B)^2\,n_1 n_l - 1 - d\,\mu_B\,n_L < 0.$$

Hence, by splitting the term $d\,\mu_B\,n_L$ from the first sum and subtracting the positive quantity $\sum_{l=2}^{L}(d\,\mu_B)^2\,n_1 n_l$ from the left hand side of the above inequality, we get

$$\left(\sum_{l=2}^{L} d\,\mu_B\,n_l + \sum_{l=2}^{L-1} d\,\mu_B\,n_l\right) + (d\,\mu_B\,n_L - d\,\mu_B\,n_L) < 1,$$

and, finally, with $\sum_{l=2}^{L-1} d\,\mu_B\,n_l \geq d\,\mu_B\,n_1$,

$$\sum_{l=1}^{L} d\,\mu_B\,n_l < 1.$$

Hence, for $L \geq 3$, Condition (D.14) implies that $k\,d\,\mu_B < 1$.

Two orthonormal bases ($L = 2$). We have to show that

$$d\,\mu_B\,(n_1 + n_2) < 1.$$

For this, consider the following maximization problem:

$$\begin{aligned}
&\max\ f(n_1, n_2) := d\,\mu_B\,(n_1 + n_2) && \text{(D.15)}\\
&\text{s.t. } g_1(n_1, n_2) := \frac{d\,\mu_B\,n_2}{1 + d\,\mu_B\,n_2} - \frac{1}{2(1 + d\,\mu_B\,n_1)} + s \leq 0, && \text{(D.16)}\\
&\qquad g_2(n_1, n_2) := n_1 - n_2 \leq 0, && \text{(D.17)}\\
&\qquad g_3(n_1, n_2) := -n_1 \leq 0, && \text{(D.18)}\\
&\qquad g_4(n_1, n_2) := -n_2 + 1 \leq 0, && \text{(D.19)}
\end{aligned}$$

where $s > 0$ is a positive constant, that allows us to restate the strict inequality constraint (D.14) as an inequality. The second constraint ensures $n_1 \leq n_2$, while the third and fourth constraints guarantee $n_1 \geq 0$ and $n_2 \geq 1$. Note, we excluded the trivial case $n_1, n_2 = 0$.

We now state the Lagrangian function corresponding to the above maximization problem and the Karush-Kuhn-Tucker (KKT) conditions that are necessary conditions for some pair of numbers n_1 and n_2 to be maximizers (local or global ones) of the objective function [28]. The Lagrangian function is

$$\mathcal{L}(n_1, n_2, \alpha_1, \alpha_2, \alpha_3, \alpha_4) := f(n_1, n_2) + \sum_{i=1}^{4} \alpha_i g_i(n_1, n_2), \tag{D.20}$$

where the α_i are constants.

The necessary KKT conditions for the pair of numbers $\hat{n_1}$ and $\hat{n_2}$ to be local maximizers of the objective function $f(n_1, n_2)$ are:

- There exist α_i, such that

  $$\nabla f(\hat{n_1}, \hat{n_2}) + \sum_{i=1}^{4} \alpha_i \nabla g_i(\hat{n_1}, \hat{n_2}) = \mathbf{0}, \tag{D.21}$$

  where $\nabla a(n_1, n_2)$ is the gradient of the function a with respect to n_1 and n_2, i.e.,

  $$\nabla a(n_1, n_2) = \begin{bmatrix} \frac{\partial a}{\partial n_1} \\ \frac{\partial a}{\partial n_2} \end{bmatrix}.$$

- $g_i(\hat{n_1}, \hat{n_2}) \leq 0$ for $i = 1, \ldots, 4$, i.e., the constraints of the maximization problem are fulfilled.

- $\alpha_i g_i(\hat{n_1}, \hat{n_2}) = 0$ for $i = 1, \ldots, 4$.

- $\alpha_i \leq 0$, for $i = 1, \ldots, 4$.

By calculating the gradient of the Lagrangian function, (D.21) becomes

$$\begin{bmatrix} d\,\mu_B \\ d\,\mu_B \end{bmatrix} + \alpha_1 \begin{bmatrix} \frac{d\,\mu_B}{2(1+d\,\mu_B\,n_1)^2} \\ \frac{d\,\mu_B}{(1+d\,\mu_B\,n_2)^2} \end{bmatrix} + \alpha_2 \begin{bmatrix} 1 \\ -1 \end{bmatrix} + \alpha_3 \begin{bmatrix} -1 \\ 0 \end{bmatrix} + \alpha_4 \begin{bmatrix} 0 \\ -1 \end{bmatrix} = \mathbf{0}. \tag{D.22}$$

We now search for all local maximizers of the maximization problem by identifying points that satisfy the KKT conditions and show that the objective function is smaller than 1. Thus, the global maximizer is also less than 1. For this we consider the settings

$$n_1 = 0, n_2 \geq 1, \text{ and}$$
$$n_1, n_2 > 0,$$

that cover all possible values of n_1 and n_2, case by case.

Case 1: $n_1 = 0, n_2 \geq 1$. In this case we do not need the KKT conditions. Observe that (D.14) implies

$$\frac{d\,\mu_B\,n_2}{1 + d\,\mu_B\,n_2} < \frac{1}{2\,(1 + d\,\mu_B\,n_1)}.$$

Hence, with $n_1 = 0$, we have

$$\frac{d\,\mu_B\,n_2}{1 + d\,\mu_B\,n_2} < \frac{1}{2}$$

Thus

$$d\,\mu_B\,n_2 < \frac{1}{2} + \frac{1}{2}\,d\,\mu_B\,n_2,$$

and hence

$$d\,\mu_B\,n_2 < 1.$$

This proves that for $n_1 = 0$ the objective function $d\,\mu_B\,(n_1 + n_2)$ is bounded above by 1.

Case 2: $n_1, n_2 > 0$. We immediately get $\alpha_3 = 0$ from $\alpha_3 g_3(\hat{n_1}, \hat{n_2}) = 0$ and $\alpha_4 = 0$ or $\hat{n_2} = 1$ from $\alpha_4 g_4(\hat{n_1}, \hat{n_2}) = 0$.

- $\alpha_4 = 0$. From $\alpha_2 g_2(\hat{n_1}, \hat{n_2}) = 0$ we get that either $\alpha_2 = 0$ or $\hat{n_1} = \hat{n_2}$.
  - $\alpha_2 = 0$. In this case Equation (D.22) becomes

    $$\begin{bmatrix} d\,\mu_B \\ d\,\mu_B \end{bmatrix} + \alpha_1 \begin{bmatrix} \frac{d\,\mu_B}{2\,(1+d\,\mu_B\,n_1)^2} \\ \frac{d\,\mu_B}{(1+d\,\mu_B\,n_2)^2} \end{bmatrix} = \mathbf{0}.$$

    This can not hold for any α_1 and hence we cannot find a maximum here.

- $\hat{n_1} = \hat{n_2} = \hat{n}$. From $\alpha_1 g_1(\hat{n_1}, \hat{n_2}) = 0$ we get

$$\begin{aligned}
\alpha_1 \left(\frac{d\,\mu_B\,\hat{n}}{1 + d\,\mu_B\,\hat{n}} - \frac{1}{2\,(1 + d\,\mu_B\,\hat{n})} + s \right) &= 0 && \iff \\
\alpha_1 \left(\frac{2d\,\mu_B\,\hat{n}}{2(1 + d\,\mu_B\,\hat{n})} - \frac{1}{2\,(1 + d\,\mu_B\,\hat{n})} + s \right) &= 0 && \iff \\
\alpha_1 \left(2\,d\,\mu_B\,\hat{n} - 1 + s' \right) &= 0,
\end{aligned}$$

where $s' = 2\,s\,(1 + d\,\mu_B\,\hat{n})$ and is thus positive.
We can deduce that either $\alpha_1 = 0$ or

$$\hat{n} = \frac{1 - s'}{2\,d\,\mu_B}.$$

In the case that $\alpha_1 = 0$, (D.22) becomes

$$\begin{bmatrix} d\,\mu_B \\ d\,\mu_B \end{bmatrix} + \alpha_2 \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \mathbf{0},$$

which cannot hold—hence in this case we will not find a maximum. In the other case, i.e., $\hat{n} = (1 - s')/(2\,d\,\mu_B)$, we have

$$\begin{bmatrix} d\,\mu_B \\ d\,\mu_B \end{bmatrix} + \alpha_1 \begin{bmatrix} \frac{d\,\mu_B}{2\,(1+d\,\mu_B\,n)^2} \\ \frac{d\,\mu_B}{(1+d\,\mu_B\,n)^2} \end{bmatrix} + \alpha_2 \begin{bmatrix} 1 \\ -1 \end{bmatrix} = \mathbf{0}.$$

The above equation holds for

$$\begin{aligned}
\alpha_1 &= -\frac{4}{3}\,(1 + d\,\mu_B\,n)^2, \text{ and} \\
\alpha_2 &= -\frac{1}{3}\,d\,\mu_B.
\end{aligned}$$

Observe that $g_i(\hat{n}, \hat{n}) = 0$ for $i = 1, \ldots, 4$, and thus all the necessary KKT conditions for a local maximum are satisfied at the point given by $\hat{n_1} = \hat{n_2} = \hat{n}$. The objective value at this point is

$$\begin{aligned}
f(\hat{n}, \hat{n}) &= d\,\mu_B \left[2\,\frac{1 - s'}{2\,d\,\mu_B} \right] \\
&= 1 - s' \\
&< 1.
\end{aligned}$$

- $\hat{n_2} = 1$. We already considered the case $\hat{n_1} = 0$ and $\hat{n_2} = 1$ before, so we now assume that $\hat{n_1} = 1$. Consider the constraint $g_1(\hat{n_1}, \hat{n_2}) = g_1(1, 1) < 0$ and follow the calculation:

$$\begin{aligned}
\frac{d\,\mu_B}{1 + d\,\mu_B} - \frac{1}{2(1 + d\,\mu_B)} + s &\le 0 && \iff \\
2\,d\,\mu_B - 1 + s' &\le 0 && \iff \\
2\,d\,\mu_B &\le 1 - s',
\end{aligned}$$

where $s' = 2\,s\,(1{+}d\,\mu_B)$ is a positive number. Observe that the left hand side of the above equation is the objective function for the setting $\hat{n_1} = \hat{n_2} = 1$. Since s' is positive we can only get $f(1,1) < 1$.

We have shown that all maxima of the objective function in (D.15) are smaller than 1. This implies $k\,d\,\mu_B < 1$ for $L = 2$. □