

Volume 7

# Academic Press Library in Signal Processing

Array, Radar and Communications Engineering

Rama Chellappa  
Sergios Theodoridis



# Academic Press Library in Signal Processing, Volume 7

Array, Radar and  
Communications Engineering

# Academic Press Library in Signal Processing, Volume 7

## Array, Radar and Communications Engineering

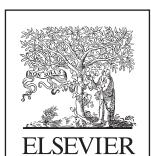
**Edited by**

**Rama Chellappa**

*Department of Electrical and Computer Engineering  
and Center for Automation Research, University of  
Maryland, College Park, MD, USA*

**Sergios Theodoridis**

*Department of Informatics & Telecommunications,  
University of Athens, Greece*



**ACADEMIC PRESS**

An imprint of Elsevier

Academic Press is an imprint of Elsevier  
125 London Wall, London EC2Y 5AS, United Kingdom  
525 B Street, Suite 1800, San Diego, CA 92101-4495, United States  
50 Hampshire Street, 5th Floor, Cambridge, MA 02139, United States  
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, United Kingdom

© 2018 Elsevier Ltd. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Details on how to seek permission, further information about the Publisher's permissions policies and our arrangements with organizations such as the Copyright Clearance Center and the Copyright Licensing Agency, can be found at our website: [www.elsevier.com/permissions](http://www.elsevier.com/permissions).

This book and the individual contributions contained in it are protected under copyright by the Publisher (other than as may be noted herein).

#### Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

#### Library of Congress Cataloging-in-Publication Data

A catalog record for this book is available from the Library of Congress

#### British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library

ISBN 978-0-12-811887-0

For information on all Academic Press publications  
visit our website at <https://www.elsevier.com/books-and-journals>



Working together  
to grow libraries in  
developing countries

[www.elsevier.com](http://www.elsevier.com) • [www.bookaid.org](http://www.bookaid.org)

*Publisher:* Mara Conner

*Acquisition Editor:* Tim Pitts

*Editorial Project Manager:* Charlotte Kent

*Production Project Manager:* Sujatha Thirugnana Sambandam

*Cover Designer:* Mark Rogers

Typeset by SPi Global, India

# Contributors

**Akram Al-Hourani**

RMIT University, Melbourne, VIC, Australia

**Shannon D. Blunt**

University of Kansas, Lawrence, KS, United States

**Shaun R. Doughty**

Dept. of Electrical and Electronic Engineering, University College London (UCL),  
Torrington Place, London, United Kingdom; Current affiliation: Maxeler  
Technologies, London, United Kingdom

**Robin J. Evans**

University of Melbourne, Melbourne, VIC, Australia

**Peter M. Farrell**

University of Melbourne, Melbourne, VIC, Australia

**Stefano Fortunati**

University of Pisa, Pisa, Italy

**Fulvio Gini**

University of Pisa, Pisa, Italy

**Nathan A. Goodman**

The University of Oklahoma, Norman, OK, United States

**Maria S. Greco**

University of Pisa, Pisa, Italy

**Yusuke Hioka**

University of Auckland, Auckland, New Zealand

**Michael R. Inggs**

Radar Remote Sensing Group (RRSG), Dept. of Electrical Engineering, University  
of Cape Town (UCT), Cape Town, South Africa

**John Jakabosky**

University of Kansas, Lawrence, KS, United States; US Naval Research  
Laboratory, Washington, DC, United States

**Kevin J. Sangston**

Georgia Tech Research Institute, Atlanta, GA, United States

**Yindi Jing**

University of Alberta, Edmonton, AB, Canada

**Sithamparanathan Kandeepan**

University of Melbourne, Melbourne, VIC, Australia

**Muhammad R.A. Khandaker**

University College London, London, United Kingdom

**Andy W.H. Khong**

Nanyang Technological University, Singapore

**Jian Li**

University of Florida, Gainesville, FL, USA

**Liang Liu**

University of Toronto, Toronto, ON, Canada

**Rohith Mars**

Nanyang Technological University, Singapore

**Marco Martorella**

University of Pisa, Pisa, Italy

**Patrick McCormick**

University of Kansas, Lawrence, KS, United States

**Justin G. Metcalf**

US Air Force Research Laboratory, Wright-Patterson AFB, Dayton, OH, United States

**Bill Moran**

RMIT University, Melbourne, VIC, Australia

**Kenta Niwa**

NTT Media Intelligence Laboratories, Tokyo, Japan

**Daniel W. O'Hagan**

Radar Remote Sensing Group (RRSG), Dept. of Electrical Engineering, University of Cape Town (UCT), Cape Town, South Africa

**Udaya Parampalli**

University of Melbourne, Melbourne, VIC, Australia

**Vaninirappuputhenpurayil Gopalan Reju**

Nanyang Technological University, Singapore

**Shahram ShahbazPanahi**

University of Ontario Institute of Technology, Oshawa, ON, Canada

**Stan Skafidas**

University of Melbourne, Melbourne, VIC, Australia

**Petre Stoica**

Uppsala University, Uppsala, Sweden

**Peng Seng Tan**

University of Kansas, Lawrence, KS, United States

**Matthias Weiß**

Fraunhofer FHR, Passive Radar and Anti-Jamming Techniques (PSR), Wachtberg, Germany

**Kai-Kit Wong**

University College London, London, United Kingdom

**Lihua Xie**

Nanyang Technological University, Singapore, Singapore

**Jie Xu**

Guangdong University of Technology, Guangzhou, Guangdong, China

**Zai Yang**

Nanjing University of Science and Technology, Nanjing, China; Nanyang Technological University, Singapore, Singapore

**Rui Zhang**

National University of Singapore, Singapore, Singapore

# About the Editors

**Prof. Rama Chellappa** is a distinguished university professor, a Minta Martin Professor in Engineering and chair of the Department of Electrical and Computer Engineering at the University of Maryland, College Park, MD. He received his BE (Hons.) degree in Electronics and Communication Engineering from the University of Madras, India and the ME (with distinction) degree from the Indian Institute of Science, Bangalore, India. He received his MSEE and PhD degrees in Electrical Engineering from Purdue University, West Lafayette, IN. At UMD, he was an affiliate professor of Computer Science Department, Applied Mathematics, and Scientific Computing Program, a member of the Center for Automation Research and a permanent member of the Institute for Advanced Computer Studies. His current research interests span many areas in image processing, computer vision, and machine learning.



Prof. Chellappa is a recipient of an NSF Presidential Young Investigator Award and four IBM Faculty Development Awards. He received the K.S. Fu Prize from the International Association of Pattern Recognition (IAPR). He is a recipient of the Society, Technical Achievement, and Meritorious Service Awards from the IEEE Signal Processing Society. He also received the Technical Achievement and Meritorious Service Awards from the IEEE Computer Society. Recently, he received the Inaugural Leadership Award from the IEEE Biometrics Council. At UMD, he received numerous college- and university-level recognitions for research, teaching, innovation, and mentoring of undergraduate students. In 2010, he was recognized as an Outstanding ECE by Purdue University. He received the Distinguished Alumni Award from the Indian Institute of Science in 2016. He is a fellow of IEEE, IAPR, OSA, AAAS, ACM, and AAAI and holds six patents to his credit.

Prof. Chellappa served the EIC of IEEE Transactions on Pattern Analysis and Machine Intelligence, as the co-EIC of Graphical Models and Image Processing, as an associate editor of four IEEE Transactions, as a co-guest editor of many special issues, and is currently on the Editorial Board of SIAM Journal of Imaging Science and Image and Vision Computing. He has also served as the general and technical program chair/co-chair for several IEEE International and National Conferences and Workshops. He is a golden core member of the IEEE Computer Society, served as a distinguished lecturer of the IEEE Signal Processing Society and as the president of IEEE Biometrics Council.

**Sergios Theodoridis** is currently professor of Signal Processing and Machine Learning in the Department of Informatics and Telecommunications of the University of Athens. His research interests lie in the areas of Adaptive Algorithms, Distributed and Sparsity—Aware Learning, Machine Learning and Pattern Recognition, Signal Processing for Audio Processing, and Retrieval.

He is the author of the book “Machine Learning: A Bayesian and Optimization Perspective,” Academic Press, 2015, the co-author of the best-selling book “Pattern Recognition,” Academic Press, 4th ed., 2009, the co-author of the book “Introduction to Pattern Recognition: A MATLAB Approach,” Academic Press, 2010, the co-editor of the book “Efficient Algorithms for Signal Processing and System Identification”, Prentice-Hall 1993, and the co-author of three books in Greek, two of them for the Greek Open University.



He currently serves as editor-in-chief for the IEEE Transactions on Signal Processing. He is editor-in-chief for the Signal Processing Book Series, Academic Press and co-editor-in-chief for the E-Reference Signal Processing, Elsevier.

He is the co-author of seven papers that have received *Best Paper Awards* including the 2014 IEEE Signal Processing Magazine best paper award and the 2009 IEEE Computational Intelligence Society Transactions on Neural Networks Outstanding Paper Award.

He is the recipient of the 2017 EURASIP *Athanasios Papoulis Award*, the 2014 IEEE Signal Processing Society *Education Award* and the 2014 EURASIP *Meritorious Service Award*. He has served as a *Distinguished Lecturer* for the IEEE Signal Processing as well as the Circuits and Systems Societies. He was *Otto Monstead Guest Professor*, Technical University of Denmark, 2012, and holder of the *Excellence Chair*, Department of Signal Processing and Communications, University Carlos III, Madrid, Spain, 2011.

He has served as president of the European Association for Signal Processing (EURASIP), as a member of the Board of Governors for the IEEE CAS Society, as a member of the Board of Governors (Member-at-Large) of the IEEE SP Society and as a Chair of the Signal Processing Theory and Methods (SPTM) technical committee of IEEE SPS.

He is a fellow of IET, a corresponding fellow of the Royal Society of Edinburgh (RSE), a fellow of EURASIP and a fellow of IEEE.

# Section Editors

## Section 1

**Fulvio Gini** (Fellow IEEE) received the Doctor Engineer (cum laude) and the Research Doctor degrees in electronic engineering from the University of Pisa, Italy, in 1990 and 1995, respectively. In 1993, he joined the Department of Ingegneriadell'Informazione of the University of Pisa, where he became an Associate Professor in 2000 and he is full professor since 2006. Prof. Gini is the Deputy Head of the Department since November 2016. From July 1996 to January 1997, he was a visiting researcher at the Department of Electrical Engineering, University of Virginia, Charlottesville. He is an Associate Editor for the IEEE Transactions on Aerospace and Electronic Systems since January 2007 and for the Elsevier Signal Processing journal since December 2006. He has been AE for the Transactions on Signal Processing (2000–06) and is a Senior AE of the same Transaction since February 2016. He was a member of the EURASIP JASP Editorial Board. He was co-founder and first co-Editor-in-Chief of the Hindawi International Journal on Navigation and Observation (2007–11). He was the area editor for the special issues of the IEEE Signal Processing Magazine (2012–14). He was a co-recipient of the 2001 and 2012 IEEE AES Society's Barry Carlton Award for Best Paper published in the IEEE Transactions on AES. He was a recipient of the 2003 IEE Achievement Award for outstanding contribution in signal processing and of the 2003 IEEE AES Society Nathanson Award to the Young Engineer of the Year. He is a member of the Radar System Panel (2008–present) and also a member of the Board of Governors (BoG) (2017–19) of the IEEE Aerospace and Electronic Systems Society (AESS). He is a member of the IEEE SPS Awards Board (2016–18). He has been a member of the Signal Processing Theory and Methods (SPTM) Technical Committee (TC) of the IEEE Signal Processing Society and of the Sensor Array and Multichannel (SAM) TC for many years. He is a member of the Board of Directors (BoD) of the EURASIP Society, the Award Chair (2006–12), and the EURASIP President (2013–16). He is the General co-Chair of the 2020 IEEE Radar Conference to be held in Florence in September 2020. He was the Technical co-Chair of the 2006 EURASIP Signal and Image Processing Conference (EUSIPCO 2006), Florence (Italy), of the 2008 Radar Conference, Rome (Italy), and of the 2015 IEEE CAMSAP workshop, Cancun (Mexico). He was the General co-Chair of the 2nd Workshop on Cognitive Information Processing (CIP2010), of the 2014 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2014), held in



Florence (Italy), and of the 2nd, 3rd, and 4th editions of the workshop on Compressive Sensing in Radar (CoSeRa). Prof. Gini was the section editor for the “Radar Signal Processing” section, Vol. 3 of the *Academic Press Library in Signal Processing*, S. Theodoridis and R. Chellappa, editors, Elsevier Ltd, 2013. He was the guest co-editor of two special sections of the Journal of the IEEE SP Society on Special Topics in Signal Processing, one on “Adaptive Waveform Design for Agile Sensing and Communication” (2007) and the other on “Advanced Signal Processing for Time/Frequency Modulated Arrays” (2017), guest editor of the special section of the IEEE Signal Processing Magazine on “Knowledge Based Systems for Adaptive Radar Detection, Tracking and Classification” (2006), guest co-editor of the two special issues of the EURASIP Signal Processing journal on “New trends and findings in antenna array processing for radar” (2004) and on “Advances in Sensor Array Processing (in memory of Alex Gershman)” (2013). He is co-editor and author of the book “Knowledge Based Radar Detection, Tracking and Classification” (2008) and of the book “Waveform Diversity and Design” (2012). He authored or co-authored 11 book chapters, about 125 journal papers, and 160 conference papers.

## Section 2

**Nikos Sidiropoulos** (Fellow, IEEE) received his PhD in 1992 from the University of Maryland, College Park, where he was affiliated with the Institute for Systems Research. He has served on the faculty of the University of Virginia, TU Crete, Greece, and University of Minnesota, where he has been a Professor of ECE since 2011, and currently holds an ADC Chair in digital technology. His research spans topics in signal processing systems theory and algorithms, optimization, communications, and factor analysis—with a long-term interest in tensor decomposition and its applications. His current focus is primarily on signal and tensor analytics for learning from big data. He received the NSF/CAREER award in 1998, and the IEEE Signal Processing Society (SPS) Best Paper Award in 2001, 2007, and 2011. He served as IEEE SPS Distinguished Lecturer (2008–09), and as Chair of the IEEE Signal Processing for Communications and Networking Technical Committee (2007–08). He served as Associate Editor for IEEE Transactions on Signal Processing (2000–06), IEEE Signal Processing Letters (2000–02), Signal Processing (2009–13), on the editorial board of IEEE Signal Processing Magazine (2009–11), and as Area Editor for IEEE Transactions on Signal Processing (2012–14). He currently serves as VP-Membership of IEEE SPS. He received the 2010 IEEE Signal Processing Society Meritorious Service Award, the 2013 Distinguished Alumni Award from the Dept. of ECE, University of Maryland, and was elected Fellow of the European Association for Signal Processing (EURASIP) in 2014.



## Section 3

**Marius Pesavento** received his Dipl.-Ing. and MEng. degrees from Ruhr-University Bochum, Bochum, Germany, and McMaster University, Hamilton, Ontario, Canada, in 1999 and 2000, respectively, and the Dr-Ing. degree in electrical engineering from Ruhr-University Bochum in 2005. Between 2005 and 2009, he held research positions in two start-up companies in the ICT area. In 2010, he became an Assistant Professor for Robust Signal Processing and a full professor for Communication Systems in 2013 at the Department of Electrical Engineering and Information Technology, Technical University Darmstadt, Darmstadt, Germany. His research interests include robust signal processing and adaptive beamforming, high-resolution sensor array processing, multiantenna and multiuser communication systems, distributed, sparse, and mixed-integer optimization techniques for signal processing, communications and machine learning, statistical signal processing, spectral analysis, and parameter estimation. He has received the 2003 ITG/VDE Best Paper Award, the 2005 Young Author Best Paper Award of the IEEE Transactions on Signal Processing, and the 2010 Best Paper Award of the CrownCOM conference. He is a member of the Editorial Board of the EURASIP Signal Processing Journal, and served as an Associate Editor for the IEEE Transactions on Signal Processing in 2012–16. He is a member of the Sensor Array and Multichannel (SAM) Technical Committee of the IEEE Signal Processing Society, and the Special Area Teams “Signal Processing for Communications and Networking” and “Signal Processing for Multisensor Systems” of the EURASIP.



## Section 4

**Patrick Naylor** is a member of academic staff in the Department of Electrical and Electronic Engineering at Imperial College London. He received the BEng degree in Electronic and Electrical Engineering from the University of Sheffield, United Kingdom, and the PhD degree from Imperial College London, United Kingdom. His research interests are in the areas of speech, audio, and acoustic signal processing. He has worked in particular on adaptive signal processing for speech dereverberation, blind multichannel system identification and equalization, acoustic echo control, speech quality estimation and classification, single- and multichannel speech



enhancement, and speech production modeling with particular focus on the analysis of the voice source signal. In addition to his academic research, he enjoys several fruitful links with industry in the United Kingdom, the United States, and Europe. He is the past-Chair of the IEEE Signal Processing Society Technical Committee on Audio and Acoustic Signal Processing and a Director of the European Association for Signal Processing (EURASIP). He has served as an Associate Editor of IEEE Signal Processing Letters and is currently a senior area editor of IEEE Transactions on Audio Speech and Language Processing.

# Introduction

Following the success of the first edition of the Signal Processing e-reference project, which was very well received by the signal processing community, we are pleased to present the second edition. Our effort in this second phase of the project was to fill in some remaining gaps from the first edition, but mainly to be currently taking into account recent advancements in the general areas of signal, image and video processing, and analytics.

The last 5 years, although in a historical perspective appear to be a short period, in the context of science, engineering, and technology were very dense in terms of results and ideas. The availability of massive data, which we refer to as Big Data, together with advances in Machine Learning and affordable GPUs, has opened up new areas and opportunities. In particular, the application of deep learning networks to problems such as face/object detection, object recognition, face verification/recognition has demonstrated superior performance that was not thought possible just a few years back. We are at a time when “caffe,” the software for implementing deep networks, is probably used more than FFT! We take comfort that the basic module in caffe is convolution, the basic building block of signal and image processing.

While one cannot argue against the impressive performance of deep learning methods for a wide variety of problems and time-tested concepts such as statistical models and inference, the role of geometry and physics will continue to be relevant and may even enhance the performance and generalizability of deep learning networks. Likewise, the power of nonlinear devices, hierarchical computing structures, and optimization strategies that are central to deep learning methodologies will inspire new solutions to many problems in signal and image processing.

The new chapters that appear in these volumes offer the readers the means to keep track of some of the changes that take place in the respective areas.

We would like to thank the associate editors for their hard work in attracting top scientists to contribute chapters in very hot and timely research areas and, above all, the authors who contributed their time to write the chapters.

# Holistic radar waveform diversity

# 1

Shannon D. Blunt\*, John Jakabosky\*,†, Patrick McCormick\*, Peng Seng Tan\*,  
Justin G. Metcalf‡

*University of Kansas, Lawrence, KS, United States*\* US Naval Research Laboratory, Washington,  
DC, United States† US Air Force Research Laboratory, Wright-Patterson AFB, Dayton, OH,  
United States‡

## 1.1 INTRODUCTION

A radar waveform is a signal defined in time, frequency, space, polarization, and modulation with the purpose of eliciting desired information from the illuminated environment. The very earliest examples of such waveforms can be found in nature used by the various forms of acoustic echolocating mammals (bats, dolphins, whales) [1–9]. For example, bats employ myriad different waveforms to enable search and acquisition of prey as well as navigation. Such waveforms include constant frequency, linear frequency modulation (LFM), hyperbolic FM (HFM), multiple harmonics, and various other nonlinear FM (NLFM) forms [4,5]. Likewise, dolphins [6,7] and whales [9] use different waveforms to navigate and hunt in underwater environments in which signal propagation can be exceedingly complex. In fact, the “mediocre equipment” [8] of dolphins belies their marvelous echolocation capability.

While it is certainly instructive to consider the types of waveforms used in nature, it is important to note that the use of such waveforms is the result of millions of years of evolutionary pressure to produce these highly integrated echolocation systems that are inherently robust and in which the whole is far superior to any individual “components.” As such, it stands to reason that we should likewise consider radar system design from an holistic perspective that encompasses the electromagnetic, systems engineering, and signal processing attributes. Moreover, growing spectrum congestion and an increasingly more complex interference environment [10–12] are driving the investigation into different forms of waveform diversity [13–17], many of which require this holistic perspective to represent the waveform structure adequately.

In this chapter, we examine the design of waveforms in the physical context of a radar system that must generate them and subsequently process the resulting echoes. The attributes of the transmitter and receiver that influence the waveform are discussed and used to inform the development of waveform structures and design

strategies that are amenable to these system effects. It is demonstrated experimentally that this holistic perspective facilitates significant enhancements in sensing performance and ultimately is expected to lead to a convergence of signal processing, systems engineering, and electromagnetics for radar design.

[Section 1.2](#) summarizes radar waveforms in widespread use, the associated pulse compression operation, and the common metrics used for waveform design. [Section 1.3](#) then discusses the important issues involved with transmitting and receiving physical radar emissions. In [Section 1.4](#) waveform implementation/design approaches are considered that realize waveforms that are amenable to physical systems, including compensation for the performance degradation introduced by transmitter distortion. Finally, [Section 1.5](#) discusses the incorporation of polarization and spatial dimensions into the waveform design framework.

## 1.2 PRACTICAL RADAR WAVEFORMS AND PULSE COMPRESSION

Before examining the capabilities and practical limitations of radar waveform diversity, it is first useful to establish the fundamentals of radar waveforms and pulse compression. This section briefly summarizes the more common waveform classes, the metrics by which they are generally evaluated and designed, and defines the structure of the waveform-generated received signal.

### 1.2.1 RADAR WAVEFORMS

From the standpoint of practical generation by a radar transmitter, frequency modulated (FM) waveforms [\[18\]](#) are attractive because they inherently possess constant amplitude and are well contained spectrally. The complex baseband representation of an arbitrary FM waveform of pulselength  $T$  (normalized to unit energy) is

$$s_{\text{FM}}(t) = \frac{1}{\sqrt{T}} \exp(j\theta(t)), \quad (1.1)$$

where  $\theta(t)$  is the instantaneous phase, and its scaled derivative

$$\frac{1}{2\pi} \frac{d\theta(t)}{dt} = f(t) \quad (1.2)$$

is the instantaneous frequency. The most notable FM waveform is linear FM (LFM) [\[19\]](#), which is arguably the most widely used waveform in operational radar systems because it is easy to implement in hardware, is tolerant to Doppler shift, and can be used in conjunction with stretch processing on receive to enable wideband operation [\[20\]](#). Thus the LFM waveform  $s_{\text{LFM}}(t)$ , also referred to as a *chirp*, has the quadratic phase function

$$\theta_{\text{LFM}}(t) = \pm\pi Bt^2/T \quad \text{for } 0 \leq t \leq T, \quad (1.3)$$

with associated linear instantaneous frequency  $f_{\text{LFM}}(t) = \pm Bt/T$ .

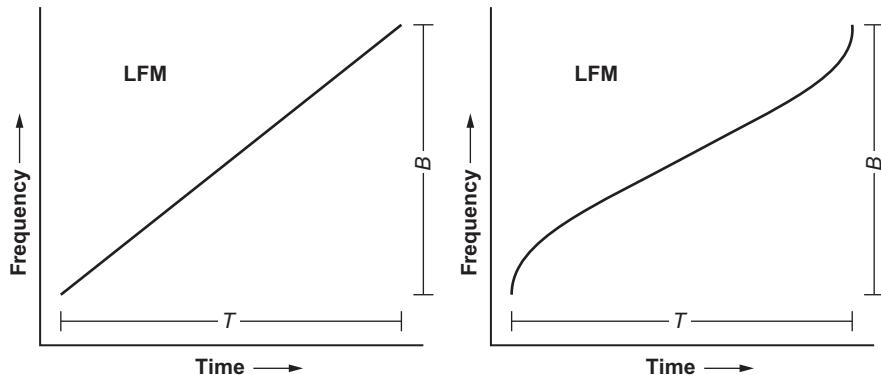
In Eq. (1.3), the term  $B$  closely approximates the 3-dB bandwidth and  $\pm$  indicates either an *up-chirp* (increasing frequency) or *down-chirp* (decreasing frequency). Further,  $BT$  denotes the *time-bandwidth product*, which indicates the SNR gain achieved when applying the matched filter.<sup>1</sup> Taking an additional derivative of Eq. (1.2) yields the instantaneous *chirp rate*

$$\frac{1}{2\pi} \frac{d^2\theta(t)}{dt^2} = \frac{df(t)}{dt} = f'(t), \quad (1.4)$$

which for LFM is the constant  $\pm B/T$ . Time-frequency plots for an LFM chirp and a generic chirp-like NLFM waveform are depicted in Fig. 1.1 for the same  $BT$ . Where the LFM has a constant chirp rate, the NLFM has a higher chirp rate at the edges with respect to the center. The purpose of the latter structure is to shape the power spectral density (PSD) in such a way as to produce lower sidelobes in the associated waveform autocorrelation (i.e., the matched filter response for zero Doppler).

For example, Fig. 1.2 illustrates the PSDs for LFM and NLFM waveforms having the same 99% power bandwidth but different 3-dB bandwidths. Since range resolution is inversely proportional to the 3-dB waveform bandwidth, it is clear that this NLFM waveform experiences some range resolution degradation relative to LFM, which is shown in Fig. 1.3. However, the more gradual spectral roll-off of NLFM is also known to produce lower range sidelobes [21], which is likewise observed in Fig. 1.3.

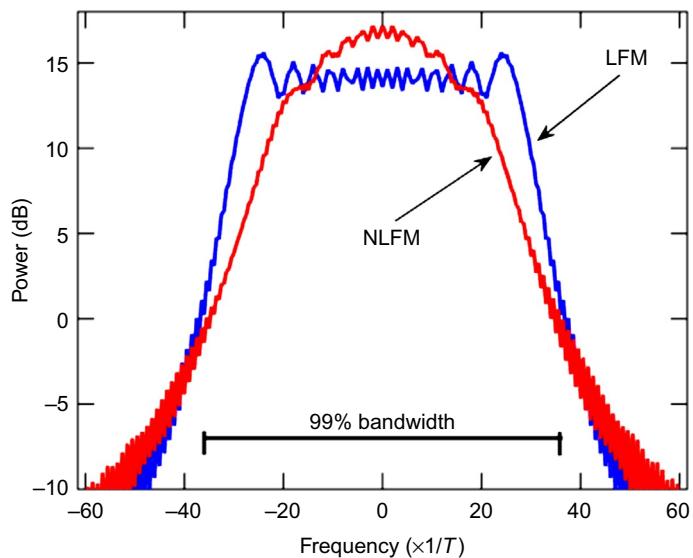
The intrinsic attributes of constant amplitude and good spectral containment makes FM waveforms naturally amenable for use with high power amplifiers (HPAs)



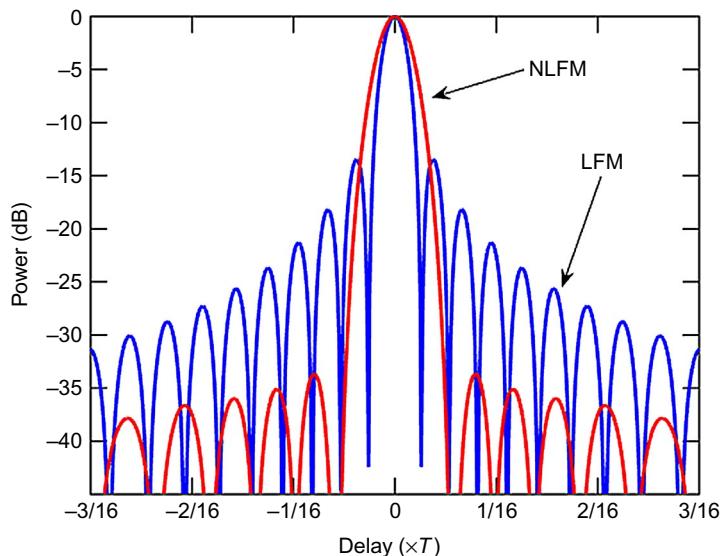
**FIG. 1.1**

Time-frequency relationship for an LFM chirp waveform (left) and a generic NLFM chirp-like waveform (right) [17].

<sup>1</sup>For arbitrary waveform  $s(t)$ , the matched filter is  $h_{MF}(t) = C s^*(T-t)$  for  $(\bullet)^*$  denoting complex conjugation and  $C$  an arbitrary constant. It is convenient to set  $C$  such that  $(\int_0^T |h_{MF}(t)|^2 dt)^{1/2} = 1$ , thus yielding a *normalized matched filter* that produces a unity noise-power gain.

**FIG. 1.2**

Power spectral densities (PSDs) of LFM and NLFM waveforms with the same 99% power bandwidth (but different 3-dB power bandwidths) [17].

**FIG. 1.3**

Matched filter responses for LFM and NLFM waveforms [17].

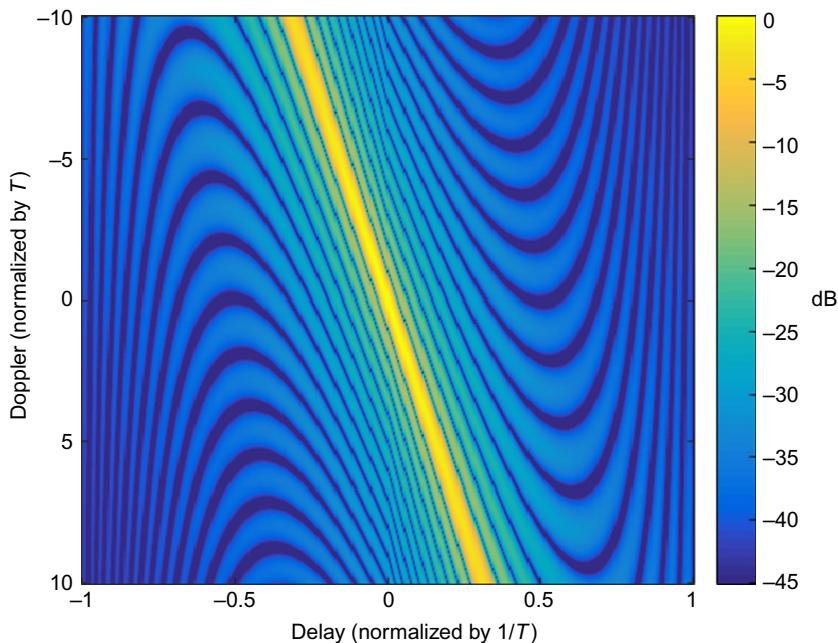
such as klystrons and traveling wave tubes (TWTs) that are widely used in many radar systems due to their high power efficiency, achievable transmit power, and reliability [22,23]. Solid-state HPAs are becoming more common as well, particularly when incorporated into an active electronically scanned array antenna architecture, though the waveform requirements are essentially unchanged.

Additional benefits of the LFM chirp, as well as sufficiently chirp-like waveforms such as in Fig. 1.1, can be observed by examining the delay-Doppler ambiguity function [18]

$$\chi(\tau, f_D) = \int_{t=0}^T e^{j2\pi f_D t} s(t) s^*(t + \tau) dt \quad (1.5)$$

that was first proposed by Woodward [24], which is essentially the matched filter response as a function of delay  $\tau$  and Doppler frequency  $f_D$ . Fig. 1.4 depicts the ambiguity function for the LFM waveform, where the LFM *Doppler tolerance* is evidenced by the delay-Doppler ridge whose peak occurs at  $(\tau=0, f_D=0)$ . Further, the important ambiguity function property ([18], Chap. 3)

$$\int_{f=-\infty}^{+\infty} \int_{\tau=-\infty}^{+\infty} |\chi(\tau, f_D)|^2 d\tau df_D = 1, \quad (1.6)$$



**FIG. 1.4**

Delay-Doppler ambiguity function for an LFM waveform [17].

assuming the waveform energy is normalized to unity with otherwise arbitrary waveform structure, represents a *conservation of ambiguity* that implies the delay-Doppler ridge for chirp and chirp-like waveforms serves to “absorb” a significant portion of the fixed amount of ambiguity. In other words, the optimization of chirp-like waveforms can generally be expected to enable much lower range sidelobes at or near zero Doppler than other waveform formulations. Examples of such chirp-like NLFM waveforms can be found in Refs. [25–30].

If the transmitter permits some degree of amplitude modulation (AM), such as through the use of solid-state HPAs or via some type of predistortion/linearization [31,32] (e.g., see Refs. [33,34]), then amplitude tapering of an FM waveform can also be performed as

$$s_{\text{Tapered-FM}}(t) = a(t)s_{\text{FM}}(t), \quad (1.7)$$

where  $0 \leq a(t) \leq 1$  for  $0 \leq t \leq T$ , as another means to reduce sidelobes ([18], Chap. 4). Such tapering is an alternative way to shape the PSD and thus, as before, can result in range resolution degradation relative to an untapered FM waveform. The tapered waveform also realizes an SNR loss compared to the untapered waveform of

$$\text{SNR Loss}_{\text{transmit taper}} = -10 \log_{10} \left[ \int_0^T a^2(t) dt \right]. \quad (1.8)$$

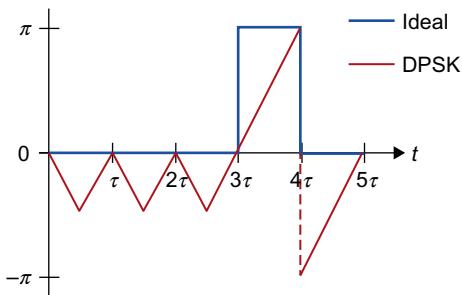
Waveforms that employ both NLFM structure and amplitude tapering are referred to as *hybrid FM* [35–37] and can generally realize very low sidelobe responses.

Besides FM, the other prominent waveform class in widespread use is that of *phase codes*, in which the pulsedwidth  $T$  is divided into a set of  $N$  subpulses, or chips, each having a duration of  $T_C = T/N$ . This phase-coded (PC) structure can be expressed as

$$s_{\text{PC}}(t) = \frac{1}{\sqrt{T}} \sum_{n=1}^N \exp(j\theta_n) \text{rect} \left[ \frac{t - (n-1)T_C}{T_C} \right] \quad \text{for } 0 \leq t \leq T, \quad (1.9)$$

where the  $n$ th chip is modulated by constant phase  $\theta_n$  that is taken from a constellation of  $P$  possible phase values. Like the general FM structure in Eq. (1.1), the energy of the PC waveform is normalized to unity. Part of the attraction of phase codes is that “good codes” can be determined by searching over the set of  $P^N$  possibilities [38,39].

The most common instantiation of phase codes in use with operational radars is the class of binary codes, for which  $P = 2$ . Well-known examples include Barker codes, minimum peak sidelobe codes, and maximal length sequences [40–43]. The common usage of binary codes is due in large part to the existence of schemes to implement them as physical waveforms that are amenable to a radar transmitter. The most well-known examples of such implementation schemes are derivative phase shift keying (DPSK) [44] and the biphase-to-quadrature transformation [45], where the latter is a form of minimum shift keying (MSK). For example,

**FIG. 1.5**

Phase trajectory of a length-5 Barker code and its DPSK implementation [17].

denoting  $s_{BC}(t)$  as the binary coded version of Eq. (1.9) via the constellation comprising  $\theta = 0$  and 180 degrees, the resulting DPSK-implemented waveform is [44]

$$s_{DPSK}(t) = s_{BC}(t - T_C/2) |\cos(\pi t/T_C)| - j s_{BC}(t) |\sin(\pi t/T_C)|, \quad (1.10)$$

which ensures the phase is continuous by avoiding the abrupt chip transitions (see Fig. 1.5).

It is the presence of these abrupt phase transitions that has previously limited the widespread usage of more general polyphase codes [46,47], which would otherwise provide greater design freedom than binary codes by virtue of  $P > 2$ . The impact of these abrupt phase transitions is discussed further in Section 1.3 and a scheme to implement arbitrary polyphase codes as FM waveforms is presented in Section 1.4. Other well-known classes of signals that have been examined for use as radar waveforms are frequency-coded signals [48–51], otherwise referred to as orthogonal frequency division multiplexing (OFDM), and noise/chaotic signals [52–56]. Both of these waveform classes inherently involve significant AM effects (or high *peak-to-average power ratio* [57,58]) that limit their usage to short-range applications due to the requirement of linear amplification to avoid distortion. Thus they have less need for a holistic perspective so we shall not consider them here.

### 1.2.2 WAVEFORM PERFORMANCE METRICS

Besides design requirements specific to a given radar mode such as bandwidth and pulsedwidth, which collectively also establish the waveform “dimensionality” via the time-bandwidth product  $BT$ , the general “goodness” of a waveform is determined according to some evaluation of the delay-Doppler ambiguity function of Eq. (1.5). Here we discuss three different metrics that arise from particular aspects of the ambiguity function.

Peak sidelobe level (PSL), or peak sidelobe ratio ([59], Chap. 20), here denoted as the operation  $\Phi_{PSL}$  on the delay-Doppler ambiguity function of Eq. (1.5), can be expressed as

$$\Phi_{\text{PSL}}[\chi(\tau, f_D = 0)] = \max_{\tau} \left| \frac{\chi(\tau, 0)}{\chi(0, 0)} \right| \quad \text{for } \tau \in [\tau_m, T] \quad (1.11)$$

for the zero Doppler cut ( $f_D = 0$ ), where the interval  $[0, \tau_m]$  corresponds to the time delay (range) mainlobe and the remaining interval  $[\tau_m, T]$  contains sidelobes. Note that  $|\chi(\tau, 0)|$  is symmetric about  $\tau = 0$ . A useful benchmark for PSL is the performance bound for the specific class of linear period modulation (LPM) waveforms [60], which are also referred to as hyperbolic FM (HFM). This bound is defined as [36]

$$\text{PSL}_{\text{LPM bound}} = [-20 \log_{10}(BT) - 3] \text{ dB} \quad (1.12)$$

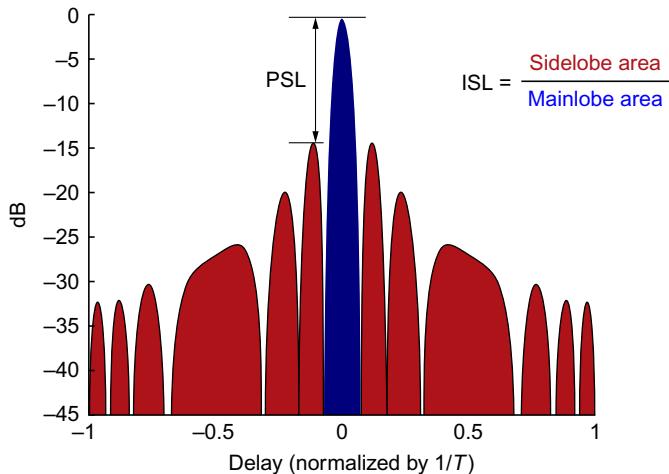
and provides a useful point of comparison for untapered FM waveforms.

Another useful metric is the integrated sidelobe level (ISL) ([59], Chap. 20) which can be defined as

$$\Phi_{\text{ISL}}[\chi(\tau, f_D = 0)] = \frac{\int_{\tau_m}^T |\chi(\tau, 0)|^2 d\tau}{\int_0^{\tau_m} |\chi(\tau, 0)|^2 d\tau} \quad (1.13)$$

for the zero Doppler cut ( $f_D = 0$ ) of the ambiguity function. Where PSL provided a measure of the largest sidelobe, ISL provides a cumulative measure of all the sidelobes and is thus useful to establish susceptibility to distributed scattering (i.e., clutter). For the  $f_D = 0$  cut, ISL and PSL are depicted conceptually in Fig. 1.6. Both of these metrics could be readily extended to also account for nonzero Doppler by establishing the mainlobe ellipse in delay-Doppler.

A third useful waveform design metric is obtained by considering the Fourier relationship between the waveform autocorrelation ( $f_D = 0$  cut of the delay-Doppler



**FIG. 1.6**

Conceptual depiction of PSL and ISL for zero Doppler [17].

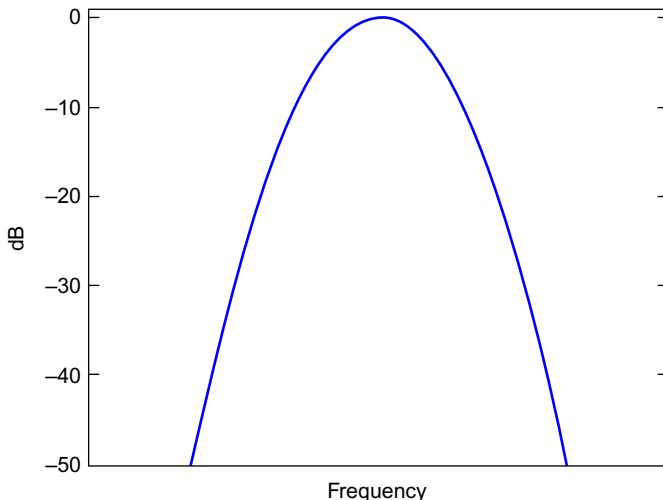
ambiguity function) and the associated PSD. We may then define a desired PSD  $|G(f)|^2$  that corresponds to a desired autocorrelation response having a sufficiently narrow mainlobe and sufficiently low range sidelobes according to some predetermined specifications. The PSL benchmark in Eq. (1.12) provides the means to determine the realism of such sidelobe goals according to the desired  $BT$ . A Gaussian-shaped PSD (Fig. 1.7) having the same energy as a constant amplitude pulse of duration  $T$  is a good example. The principle of stationary phase [26], which relates the PSD and chirp rate at a given frequency, is a well-known means with which to map a desired PSD into a NLFM waveform (see Chap. 5 of Ref. [18]). However, the desired PSD can also be used to define various optimization metrics in the frequency domain.

For instance, the frequency template error (FTE) metric defined in Ref. [29] as

$$\Phi_{\text{FTE}}[S(f), G(f)] = \left( \frac{1}{f_H - f_L} \right) \int_{f_L}^{f_H} | |S(f)|^p - |G(f)|^p |^q df, \quad (1.14)$$

provides a measure of “how close” the waveform magnitude spectrum  $|S(f)|$  is to the desired magnitude spectrum  $|G(f)|$  over the frequency interval  $f_L$  to  $f_H$ , which should include enough of the spectral roll-off to provide sufficient fidelity. The positive real values  $p$  and  $q$  define the degree of emphasis placed on different frequencies, where setting  $p = 1$  and  $q = 2$  realizes a form of frequency-domain mean-square error.

It should be noted that numerous local minima exist for these waveform design metrics. It is not necessary to determine the waveform that attains the global minimum solution as long as a given local minimum achieves some predetermined performance specifications (e.g., prescribed PSL or ISL). However, even the determination of sufficiently good local minima may be a challenge (see “performance



**FIG. 1.7**

Gaussian PSD in dB [17].

“diversity” approach in Ref. [29]). In Section 1.4 two different schemes for physical waveform design are presented. One relies upon an underlying parameterizing structure for the waveform implementation and then performs a search over the high-dimensional parameterization. The other leverages the well-known alternating projection framework for a sufficiently high fidelity representation of the waveform to ensure a physical transmitter can faithfully generate it.

### 1.2.3 RECEIVED SIGNAL STRUCTURE

Regardless of the structure of waveform  $s(t)$  or the metric used to design it, the reflected signal at the radar receiver is

$$y(t) = \int [s(t) e^{j2\pi f_D t}] * x(t, f_D) df_D + v(t), \quad (1.15)$$

where  $x(t, f_D)$  is the unknown scattering response of the illuminated environment as a function of time delay and Doppler frequency,  $v(t)$  is additive noise,  $*$  represents convolution, and the integral is taken over the possible Doppler frequencies induced by radial target/platform motion. For  $f_D = 0$  the matched filter estimate of the unknown scattering is

$$\hat{x}_{\text{MF}}(t, f_D = 0) = h_{\text{MF}}(t) * y(t). \quad (1.16)$$

The matched filter responses for other Doppler frequencies can be obtained by frequency shifting the filter  $h_{\text{MF}}(t)$  accordingly. It has also become increasingly more common to perform this pulse compression operation in the digital domain, which is discussed further in Section 1.3. Before discussing physical waveform optimization and subsequent higher dimensional extension, it is instructive to also consider the practical aspects of radar pulse compression.

## 1.3 PRACTICAL CONSIDERATIONS

There has been extensive research on waveform design, receive processing, and myriad different waveform diversity approaches [13–15,17,61], the majority of which has been largely theoretical in nature. However, there are important physical attributes of both the radar transmitter and the received echoes that must be considered if such theoretical developments are to be transitioned into practice.

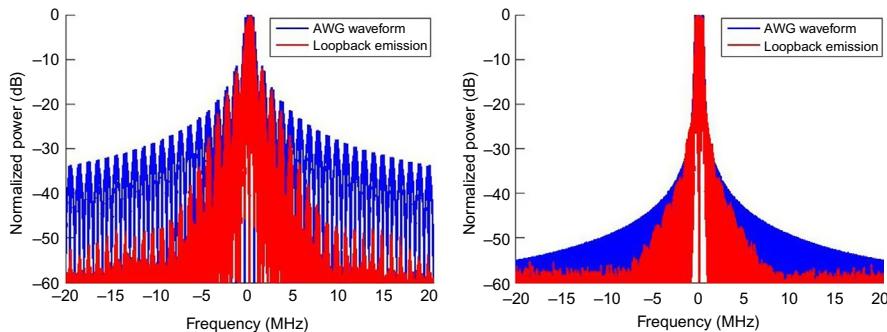
### 1.3.1 TRANSMITTER EFFECTS

The considerable time and effort that may go into the optimization of a waveform could be wasted if the impact of the transmitter is not adequately considered. The purpose of the transmitter is to generate and amplify the waveform such that the receiver can adequately capture the reflected echoes of much lower power. While legacy systems still use frequency swept local oscillators and surface acoustic wave devices for waveform generation, modern radars are moving toward the tremendous

flexibility afforded by arbitrary waveform generators (AWGs) and direct digital synthesizers [62–64]. Following waveform generation, the high-power amplifier (HPA) then produces the necessary emitted power, typical values for which could be  $\sim 100$  W up to several Megawatts [12] (note that for emerging civil applications such as automotive radar that operate over short distances the emitted power could be much less).

The overall transmit chain introduces both linear and nonlinear distortion to the intended waveform. Linear distortion results from the inherent spectral shaping of the individual transmitter components, producing amplitude ripple and phase distortion (dispersion). Nonlinear distortion is mainly due to the HPA operating in saturation to achieve high transmit power, thus generating intermodulation products from the pairwise multiplication of different frequency components in the waveform [65]. These intermodulation products are frequency harmonics that leak into the surrounding spectrum, an effect that is collectively known as *spectral regrowth* [12].

Transmitter distortion is arguably the primary reason why the use of polyphase codes has thus far been rather limited. The abrupt transitions between adjacent chips in the code correspond to out-of-band spectral content that is distorted by the spectral shaping of the transmitter components, producing AM effects that are further distorted by the saturated HPA. It is for this reason that binary codes are implemented via MSK or DPSK, such as via Eq. (1.10). For example, Fig. 1.8 shows the spectral content of a  $N = 64$  chip P4 code [38], which represents a complex baseband sampled version of an LFM waveform, and the spectral content of an actual LFM waveform with the same  $BT$  for comparison [47]. Using the form described in Eq. (1.9) for the coded waveform, both the P4 and LFM were implemented on an AWG and driven into an S-band radar test bed comprised of a mixer, preamplifier, bandpass filter, and a class AB solid-state Gallium Arsenide (GaN) HPA. The resulting “emissions” were captured by a receiver in a loopback configuration (i.e., not emitted into free space) and subsequently down-converted to baseband, analog lowpass filtered,



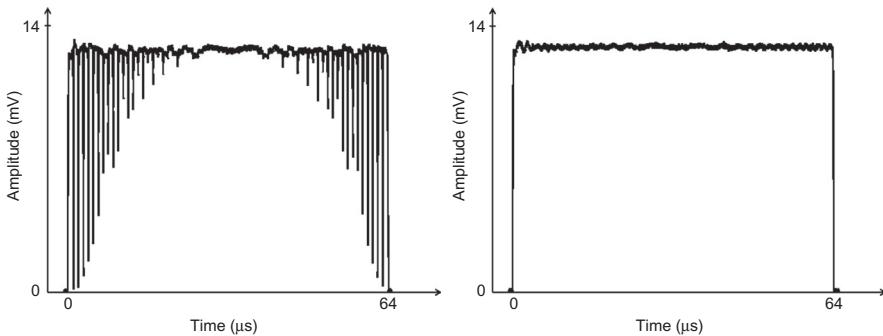
**FIG. 1.8**

Spectral content of (left) P4 code before/after transmitter distortion and (right) LFM waveform before/after transmitter distortion [47].

and then sampled at the same rate as the version of each waveform loaded onto the AWG. Clearly the transmitter spectral shaping significantly alters the extended spectral content of the P4-coded emission. In contrast, there is much less impact to the LFM waveform. Note that appreciable spectral regrowth is not observed here due to the use of a Class AB solid-state HPA, as compared to what occurs for tube-based HPAs that produce much greater output power and distortion.

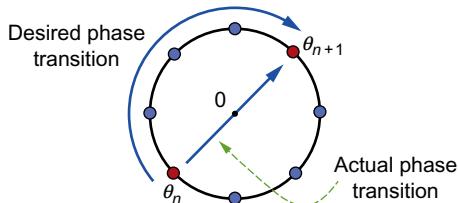
It is also instructive to examine the pulse shape of each of these waveforms after transmitter distortion. Fig. 1.9 shows that the abrupt phase transitions of the P4-coded waveform produce amplitude nulls commensurate with the amount of phase change, which for P4 are greatest near the beginning and end of the code. By comparison, the transmitter-distorted LFM only has a small amount of amplitude ripple. For a high power transmitter, these nulls correspond to power that is not transmitted and may ultimately be converted into heat that may subsequently produce increased phase noise, thereby further degrading fidelity.

Fig. 1.10 provides an illustration [17] of why coded waveforms that have not otherwise undergone some code-to-waveform (C2W) implementation (e.g., DPSK for binary codes) exhibit the response observed in Fig. 1.9 when transmitted. An FM waveform is continuous in phase and thus moves around the unit phase circle (the “desired phase transition” in the figure). However, the abrupt phase transitions



**FIG. 1.9**

Pulse shape after transmitter distortion for (left) P4 code and (right) LFM waveform [47].



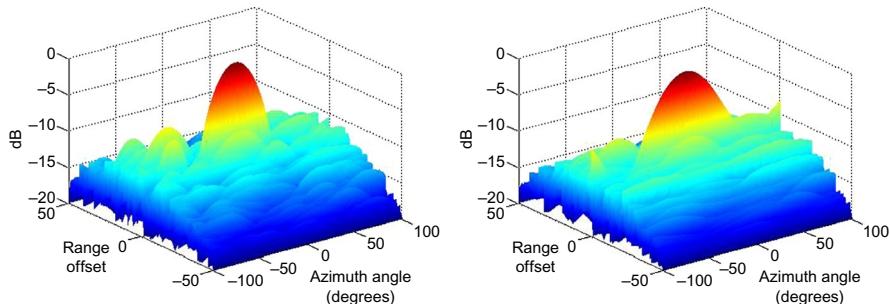
**FIG. 1.10**

Desired and actual phase transitions for a phase code due to transmitter effects [17].

involved with a coded waveform take the shortest path, which means moving through the interior of the unit circle, thereby translating into amplitude nulls. Clearly, the greater the amount of phase change, up to a maximum of 180 degrees, the deeper the null since the abrupt phase transition would come closer to the zero value at the center.

For radar modes in which different waveforms are emitted from the antenna elements in an array, otherwise known as colocated MIMO, the electromagnetic attributes of the antenna must also be considered. Antenna arrays inherently possess mutual coupling between the antenna elements, which involves neighboring elements receiving and reradiating the waveform from a given element. For an intended set of MIMO waveforms, this effect produces a distortion of the far-field delay-angle emission structure relative to an idealized case involving no mutual coupling [66,67] (Fig. 1.11).

Wideband operation presents another practical issue for MIMO emissions. Wavelength  $\lambda$  corresponding to the center frequency is an adequate approximation for narrowband operation (generally 10% bandwidth or less) to establish the antenna interelement spacing of  $d$  ( $=\lambda/2$ ). However, imaging modes such as synthetic aperture radar (SAR) generally emit wideband signals to provide fine range resolution. One could set  $d=\lambda_{\min}/2$ , for  $\lambda_{\min}$  corresponding to the highest in-band frequency to avoid grating lobes [68]. However, this choice yields interelement spacing of  $d/\lambda \ll 0.5$  for the longer wavelengths (lower frequencies), thereby resulting in “emission” into the *imaginary space* (or *invisible space*) [69] that exists beyond the endfire spatial directions at  $\phi = \pm 90$  degrees. The reality of this effect is that energy is stored in the reactive near field of the array that can lead to large amounts of power being reflected back into the transmitter, potentially damaging the radar [70]. Thus practical wideband MIMO waveform design must avoid exceeding the boundaries of real space [71,72].



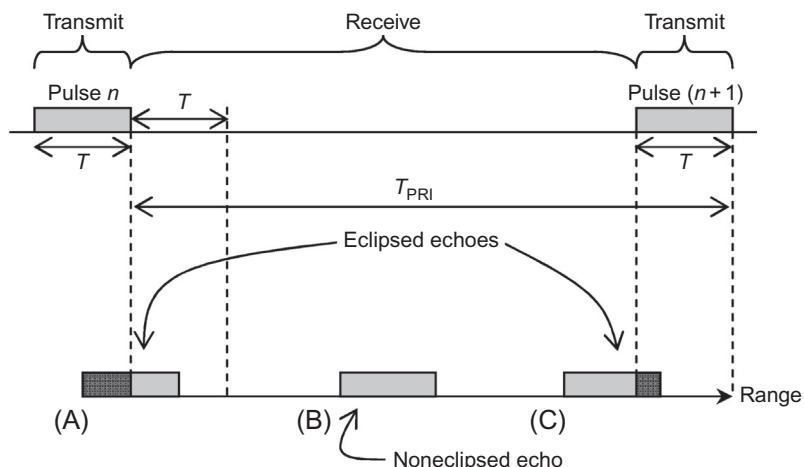
**FIG. 1.11**

Delay-angle ambiguity function for 16 waveforms generated via DPSK implementation of length-50 random binary codes where (left) no mutual coupling is present and (right)  $-10$  dB nearest neighbor mutual coupling is present but not accounted for on receive. The result is degraded resolution and  $1.1$  dB mismatch loss. [66]

### 1.3.2 RECEIVE EFFECTS

In addition to transmitter effects, the holistic perspective also necessitates consideration of several practical receive effects as well. In general, the goal of all radar modalities is to measure some desired phenomena (e.g., detect/track/image/classify targets of interest) as accurately as possible in the presence of noise and various possible forms of interference. Doing so requires that the receiver has sufficient dynamic range to capture what may be a significant power disparity among scatterer responses (perhaps several orders of magnitude) and the means with which to discriminate between signals of interest and noise/interference. The former is a driver for increasing receiver bit depth, enhanced sidelobe suppression capabilities, and possibly even migration of some interference cancellation operations back into the analog domain (e.g., DARPA program on Signal Processing at RF (SPAR) [73]). The latter (discrimination capability) is the justification for using Doppler to separate moving targets from stationary ground clutter and the subsequent need for coupled-domain formulations such as space-time adaptive processing (STAP) when the radar platform itself is moving. Enhancing discrimination capability is likewise a driver for the exploration of waveform diversity modes such as MIMO, waveform agility, and polarimetric operation to exploit the associated increase in available degrees of freedom [13–15,17,61].

While the previous separability requirements are largely a matter of system/waveform design, other factors arise because of how the radar interacts with the phenomenology of the illuminated environment. For example, to avoid damaging sensitive receiver components, pulsed radars generally turn off the receiver while the radar is transmitting. As a result, pulse eclipsing [74,75] (Fig. 1.12) can occur



**FIG. 1.12**

Echoes (A) and (C) are eclipsed because they arrive at the receiver when the radar is transmitting, while echo (B) is not eclipsed [75].

when the receiver turns on/off during the reception of a waveform-induced echo. These eclipsed echoes experience reduced SNR relative to noneclipsed echoes because a portion of the reflected pulse is not received. For chirped waveforms such as LFM, an eclipsed echo will also possess degraded range resolution since a portion of the waveform bandwidth is likewise not captured. For pulsedwidth  $T$  and pulse repetition interval  $T_{\text{PRI}}$ , the likelihood of an eclipsed echo occurrence increases as the duty cycle  $T/T_{\text{PRI}}$  is increased.

Another practical consideration arises when performing pulse compression digitally, which may necessarily be the case for many waveform-diverse operating modes. After analog antialiasing filtering and analog-to-digital (A/D) conversion, and neglecting fast-time Doppler during the pulsedwidth, the continuous baseband received signal from Eq. (1.15) can be expressed in discrete notation as

$$y(n) = \mathbf{x}^T(n)\mathbf{s} + v(n). \quad (1.17)$$

The length- $N$  vector  $\mathbf{s} = [s_1 \ s_2 \ \dots \ s_{N-1}]^T$  is the discretized version of the waveform, where  $N \approx BT$  constitutes the nominal sampling to capture the 3-dB bandwidth of the response induced by the given waveform. Likewise, the vector  $\mathbf{x}(n) = [x(n) \ x(n-1) \ \dots \ x(n-N+1)]^T$  is the collection of  $N$  contiguous samples of the unknown illuminated scattering,  $(\bullet)^T$  is the transpose operation, and  $v(n)$  is a sample of additive noise. Collecting  $N$  contiguous samples of  $y(n)$  from Eq. (1.17) to form the vector  $\mathbf{y}(n)$ , the discretized representation of the matched filter response from Eq. (1.16) is

$$\hat{x}_{\text{MF}}(n) = \mathbf{h}_{\text{MF}}^H \mathbf{y}(n) = C \mathbf{s}^H \mathbf{y}(n), \quad (1.18)$$

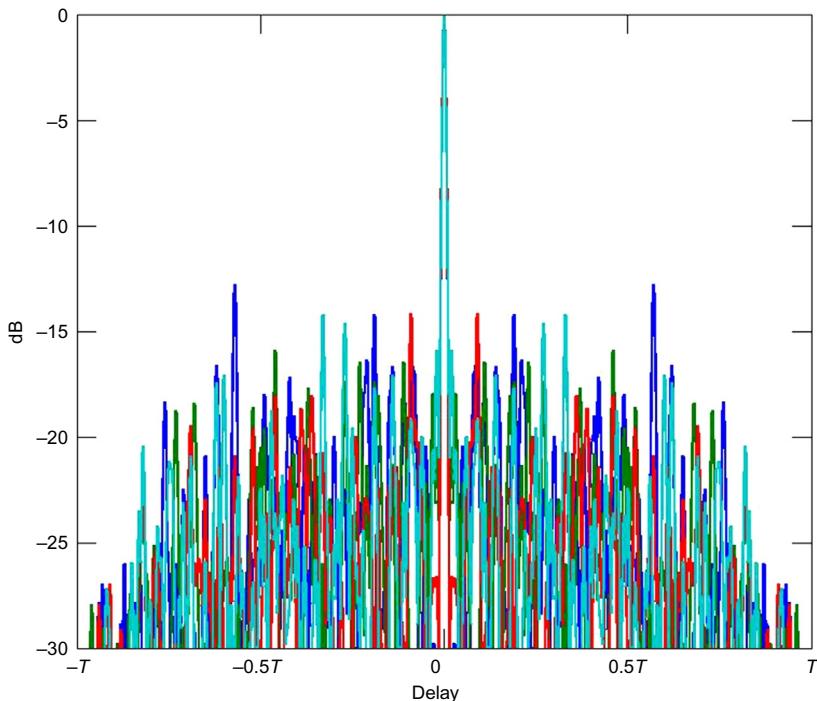
for  $(\bullet)^H$  the complex-conjugate transpose (Hermitian) operation and the scalar  $C$  again set to provide unity noise-power gain ( $\|\mathbf{h}_{\text{MF}}\| = 1$ ).

It is important to note that the finite time support of pulsed waveforms corresponds to a theoretically infinite bandwidth, thus Nyquist sampling cannot be achieved. From a practical standpoint, however, the spectral roll-off at some point falls below the noise floor, thereby establishing a finite noise-limited bandwidth which can generally be expected to be greater than the 3-dB waveform bandwidth  $B$  that is associated with range resolution. If one were to perform receive sampling at a rate commensurate with an length- $N$  discretized waveform (for  $N \approx BT$  as defined before), then the relative delay of a reflected echo could be offset by an amount that introduces a significant mismatch loss after matched filtering (up to a couple dB [76]). This effect is known as *range straddling* or *scalloping* and arises due to undersampling relative to the Nyquist rate ([59], Chap. 20).

Mitigating this range straddling effect can be achieved by simply increasing the nominal 3-dB sampling rate by some factor  $K$ . In so doing the discretized versions of the waveform and subsequent scattering become  $\mathbf{s} = [s_1 \ s_2 \ \dots \ s_{NK-1}]^T$  and  $\mathbf{x}(n) = [x(n) \ x(n-1) \ \dots \ x(n-NK+1)]^T$ , respectively, which are both now length  $NK$ . Likewise,  $NK$  contiguous samples of  $y(n)$  are now collected to form  $\mathbf{y}(n)$ , to which the length  $NK$  matched filter is applied (still normalized to produce unity noise gain). Aside from the increased computational cost to perform pulse compression at

the higher sampling rate, this modification would appear rather trivial. However, if one wishes to employ some form of optimal [77] or adaptive [78,79] pulse compression, it is also necessary to consider the mismatch losses that can arise from unintended range super-resolution [80]. Such effects can be addressed via judicious use of “beam-spoiling” in the range domain filtering [81] to realize near-nominal range resolution (i.e., the same as that of the matched filter).

Finally, in a similar manner to the delay-angle coupling achieved with MIMO, the waveform diversity concept referred to as *waveform agility* or *pulse agility* realizes a coupling in delay-Doppler (or fast time/slow time to be more precise) via the use of different waveforms over the coherent processing interval (CPI) [82–88]. By employing multiple waveforms, this operating mode could, for example, facilitate the embedding of communication information into the radar emission [82]. However, when performing pulse compression on these different waveforms the differences in their sidelobe structure induces a clutter *range sidelobe modulation* effect [82] that can impede clutter cancellation. Fig. 1.13 illustrates the matched filter responses to each of four randomly generated binary codes that are implemented with



**FIG. 1.13**

Matched filter responses for four length-100 random binary codes implemented with DPSK. The different sidelobes would modulate clutter, thereby impeding cancellation.

DPSK to produce physical waveforms. The sidelobes are clearly quite different across the set of responses, which would induce significant modulation of the clutter.

For sensing applications that rely on high dimensionality (i.e., high  $BT$  waveform and long CPI) one can expect these modulated sidelobes to simply average out and drop below the noise floor. Noise/chaotic radar [52–55] and recent work on FM noise radar [85,88] fit in this category. However, for modes that employ lower  $BT$  waveforms and a shorter CPI, and which perform clutter cancellation, it is necessary to compensate for the range sidelobe modulation effect. Such compensation can either take the form of waveform/filter optimization that serves to homogenize the sidelobe responses (in the region of zero Doppler) [82,83,86,88] or to address pulse compression and Doppler processing (slow-time) in a joint manner [83,84,87]. While the joint domain schemes have higher computational cost, the increased degrees of freedom also facilitates sufficient dimensionality to address *multiple-time-around clutter*, also known as *range-ambiguous clutter* or *folded clutter*, which is more prevalent at higher PRF ([89,90] and see Chap. 9.5 of Ref. [91]).

If only two different waveforms are used (such that each pulse represents  $\log_2(2) = 1$  bit if intended to convey information), then the sidelobe similarity constraint [82]

$$h_{\text{MF},1}(t)^*s_1(t) = h_{\text{MF},2}(t)^*s_2(t) \quad (1.19)$$

can be met by setting  $s_2(t) = s_1^*(T-t)$ , such that  $h_{\text{MF},1}(t) = s_2(t)$  and  $h_{\text{MF},2}(t) = s_1(t)$ , under the condition of negligible fast-time Doppler. However, for more than two waveforms the frequency response of the  $m$ th filter would have to be

$$H_m(\omega) = \frac{S_1(\omega)H_1(\omega)}{S_m(\omega)}, \quad (1.20)$$

for this same condition to hold. Due to the term in the denominator, the filter in Eq. (1.20) is infinite impulse response (IIR) and thus can only be approximated by a long pulse compression mismatched filter implemented as finite impulse response (FIR). Approaches to design these sidelobe-homogenizing mismatched filters have been described in Refs. [82,83,86,88].

## 1.4 HOLISTIC WAVEFORM IMPLEMENTATION AND DESIGN

In this section we consider the formulation of waveforms that inherently address the practical effects discussed previously. In so doing, the implementation and optimization of physically realizable waveforms can be achieved, thereby facilitating real-world applications of radar waveform diversity.

Specifically, two distinct optimizable waveform structures are posed that can be readily implemented on fielded radar systems. The polyphase-coded FM (PCFM) scheme provides the means to convert arbitrary polyphase codes into FM waveforms, with different variants achieving a wide assortment of physical waveforms.

In contrast to the defined structure of PCFM, a spectral shaping form of alternating projections is also discussed that, by appropriately accounting for spectral content, provides another means with which to optimize physically realizable waveforms. Finally, separate from these waveform structures, the notion of “transmitter in the loop” optimization is described whereby the distortion imposed by the transmitter onto the waveform is incorporated into the design process to realize optimization of the actual physical emission launched by the radar into free space.

### 1.4.1 POLYPHASE-CODED FM

The properties of a waveform that make it amenable to a high power radar are (1) constant amplitude and (2) sufficient spectral containment. The former reduces the impact of nonlinear distortion by avoiding the saturating effect a HPA could otherwise have upon any amplitude modulated (AM) characteristics of the waveform. Constant amplitude likewise maximizes the “energy on target,” which translates into detection sensitivity. Spectral containment reduces the impact of spectral shaping effects by the transmitter that could produce additional AM that would lead to further distortion in the HPA. Based on these properties it is not surprising that FM waveforms are an attractive choice for many radar applications.

FM radar waveforms have been in use for more than 50 years [19], with such waveforms generally possessing a “chirping” time/frequency structure such as shown in Fig. 1.1. Aside from the standard LFM, the design of such waveforms, all of which are thus nonlinear FM (NLFM), tends to rely on the determination of a suitable frequency function of time (e.g. [25–28,35–37]). In contrast, binary codes implemented with DPSK or MSK [49,50] have constant amplitude and relatively good spectral containment yet are designed via search over a parameterized code space. Because the code length  $N$  is a good approximation for the waveform time-bandwidth product  $BT$ , while traditional FM waveforms tend to be based on a relatively small number of parameters, one can argue that the coded approach makes better use of the available degrees of freedom from the standpoint of optimization. That said, it should be noted that the chirp-like time/frequency structure can support quite low range sidelobes due to the conservation of ambiguity (see Section 1.2). From the previous discussion, it can be surmised that a code-based chirp-like waveform would have considerable potential in terms of minimizing range sidelobes.

Using the continuous phase modulation framework [92], which is commonly used in aeronautical telemetry [93], deep-space communications [94], and the Bluetooth wireless standard [95], a PCFM radar waveform implementation has been developed [46,47] that converts an arbitrary polyphase code  $[\theta_0 \ \theta_1 \ \theta_2 \ \dots \ \theta_N]$  of length  $N+1$  into an FM waveform with  $BT \cong N$ . To do so, first a train of  $N$  impulses with time separation  $T_p$  is formed such that pulsedwidth  $T = NT_p$ . The  $n$ th impulse is weighted by  $-\pi \leq \alpha_n \leq \pi$ , which is the phase change occurring over time interval  $T_p$ . From a design standpoint, it is possible either to determine the  $\alpha_n$  values directly or to obtain them from an existing length  $N+1$  polyphase code via

$$\alpha_n = \begin{cases} \tilde{\alpha}_n & \text{if } |\tilde{\alpha}_n| \leq \pi \\ \tilde{\alpha}_n - 2\pi \operatorname{sgn}(\tilde{\alpha}_n) & \text{if } |\tilde{\alpha}_n| > \pi, \end{cases} \quad (1.21)$$

where

$$\tilde{\alpha}_n = \theta_n - \theta_{n-1} \quad \text{for } n = 1, \dots, N, \quad (1.22)$$

and  $\operatorname{sgn}(\bullet)$  is the sign operation.

For a phase-change code  $\mathbf{x} = [\alpha_1 \ \alpha_2 \ \dots \ \alpha_N]^T$  and arbitrary starting phase  $\theta_0$ , the physical PCFM waveform is [47]

$$s_{\text{PCFM}}(t; \mathbf{x}) = \exp \left\{ j \left( \int_0^t g(\tau)^* \left[ \sum_{n=1}^N \alpha_n \delta(\tau - (n-1)T_p) \right] d\tau + \theta_0 \right) \right\}, \quad (1.23)$$

where the shaping filter  $g(t)$  must integrate to unity over the real line and have time support on  $[0, T_p]$ , and  $*$  denotes convolution. For example, a rectangular filter scaled by  $1/T_p$  meets these requirements and, when inserted into Eq. (1.23), provides a piece-wise linear phase function that can be viewed as a first-order hold representation. In contrast, the phase-code structure of Eq. (1.9) can be viewed as a zero-order hold representation, since the phase is constant between the abrupt transitions.

Assuming the possible values of  $\alpha_n$  are drawn from a uniform discretization of the phase-change interval  $[-\pi, +\pi]$  and assigning  $\ell \in 1, 2, \dots, L$  as the indices of the set of possible phase transitions, then the  $n$ th phase transition can be defined as [47]

$$\alpha_n = 2\pi \left( \frac{\ell(n) - 1}{L - 1} \right) - \pi \quad (1.24)$$

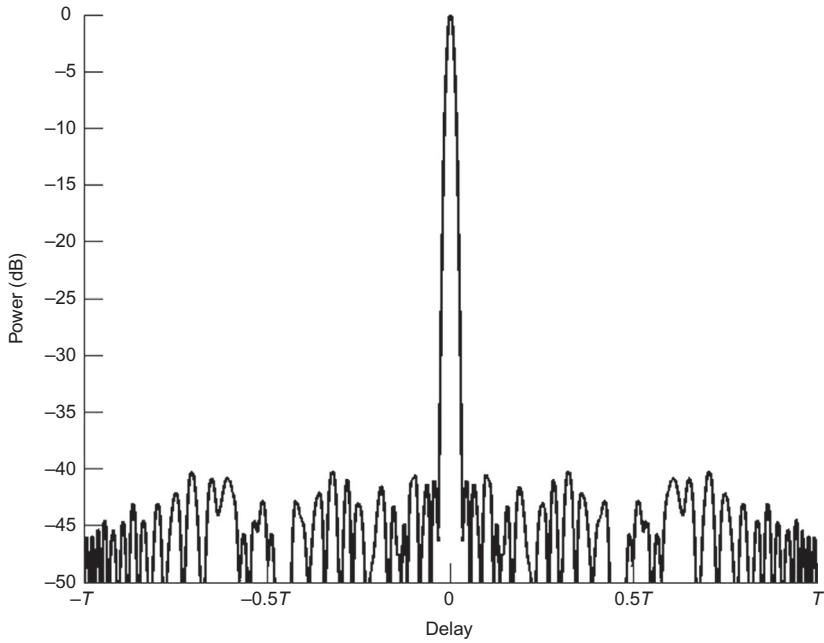
for  $n = 1, 2, \dots, N$ . Using a rectangular filter for  $g(t)$ , a piece-wise linear approximation to an LFM waveform can be achieved using Eq. (1.23) when  $\ell(n) = n$ . Design of such waveforms [29,96] involves determination of the code  $\mathbf{x} = [\alpha_1 \ \alpha_2 \ \dots \ \alpha_N]^T$  that optimizes some desired metric, such as those defined in Section 1.2. For example, Fig. 1.14 depicts an optimized PCFM waveform from Ref. [29] that has  $BT \cong N = 64$ . Compared to the LPM PSL bound from Eq. (1.12), that is found to be  $-39.1$  dB for this case, the optimized PCFM waveform realizes a PSL of  $-40.2$  dB, surpassing the bound by  $1.1$  dB.

The continuous phase component of the PCFM implementation in Eq. (1.23) can also be written as

$$\theta_{1\text{st}}(t; \mathbf{x}_1) = \int_0^t \left[ \sum_{n=1}^N \alpha_n g_1(\tau - (n-1)T_p) \right] d\tau + \theta_0, \quad (1.25)$$

where the “1” subscript in  $g_1(t)$  and  $\mathbf{x}_1 = [\alpha_1 \ \alpha_2 \ \dots \ \alpha_N]^T$  denotes this scheme as a first-order representation. Higher-order phase functions have also recently been defined in Ref. [30]. For example, second-order and third-order coded representations can be expressed as

$$\theta_{2\text{nd}}(t; \mathbf{x}_2) = \int_0^t \int_0^\tau \left[ \sum_{n=1}^N b_n g_2(\tau' - (n-1)T_p) \right] d\tau' d\tau + \int_0^t \omega_0 d\tau + \theta_0, \quad (1.26)$$

**FIG. 1.14**

Autocorrelation of an optimized PCFM waveform with  $BT = 64$  [29].

and

$$\theta_{3rd}(t; \mathbf{x}_3) = \int_0^t \int_0^\tau \int_0^{\tau'} \left[ \sum_{n=1}^N c_n g_3(\tau'' - (n-1)T_p) \right] d\tau'' d\tau' d\tau + \int_0^t \int_0^\tau \beta_0 d\tau' d\tau + \int_0^t \omega_0 d\tau + \theta_0, \quad (1.27)$$

respectively, where  $\mathbf{x}_2 = [b_1 \ b_2 \ \dots \ b_N]^T$  and  $\mathbf{x}_3 = [c_1 \ c_2 \ \dots \ c_N]^T$  are *frequency-change* (chirp rate) and *chirp-rate-change* (or “chirp acceleration”) codes, respectively, with associated shaping filters  $g_2(t)$  and  $g_3(t)$ . Also,  $\theta_0$  is the starting phase as defined for Eq. (1.23), while  $\omega_0$  and  $\beta_0$  are the initial frequency and initial chirp rate, respectively.

The continuous-phase coding structures in Eqs. (1.25), (1.26), and/or (1.27) can also be combined [30] to allow for multiorder coding to provide even greater freedom in FM waveform design, which can facilitate even lower range sidelobes near zero Doppler. As an example, Fig. 1.15 depicts the autocorrelations of a waveform obtained by joint optimization of first-order and second-order codes (combination of Eqs. 1.25 and 1.26) as well as joint optimization of first-, second-, and third-order codes (combination of Eqs. 1.25–1.27). As before, these waveforms possess

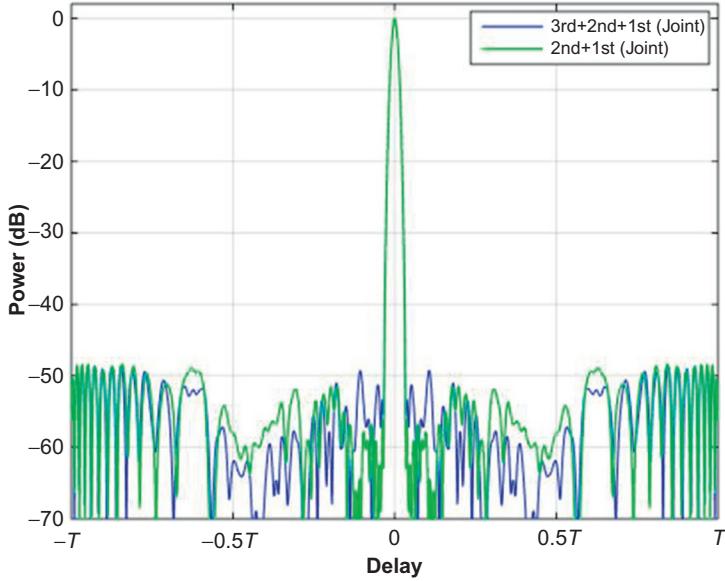


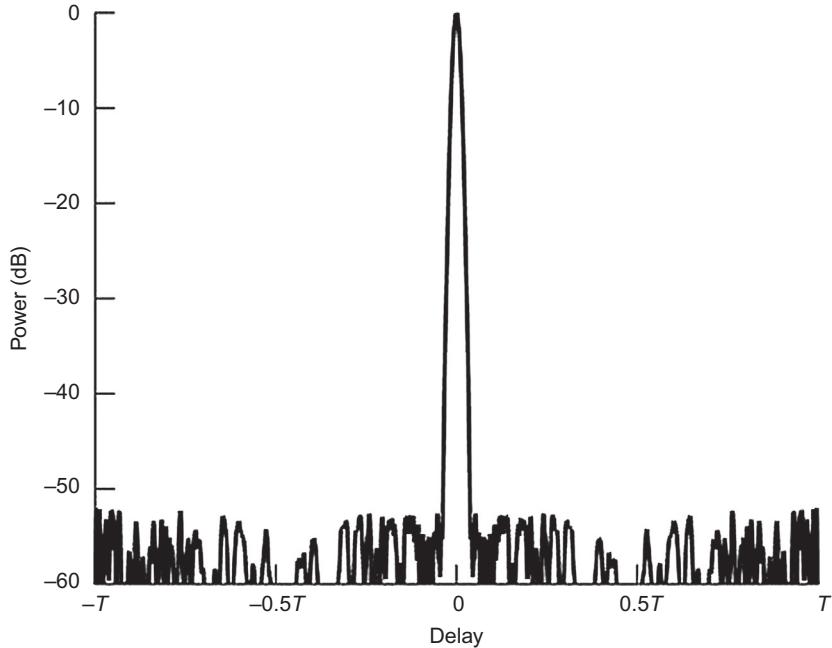
FIG. 1.15

Autocorrelation of optimized higher-order PCFM waveforms for  $BT = 64$  [30].

$BT = 64$ , where the bandwidth is constrained by imposing a Gaussian spectral mask as discussed in Section 1.2 for the FTE metric. In comparison with the associated LPM PSL bound of  $-39.1$  dB using Eq. (1.12), these waveforms realize PSL values of  $-48.4$  and  $-48.7$  dB, respectively; an improvement of more than  $9$  dB.

In addition to higher-order phase functions, the PCFM implementation of Eq. (1.23) can also be generalized to permit *overcoding* [97]. The overcoded formulation alters Eq. (1.23) in two ways: (1) the phase-change intervals of  $T_p$  are subdivided into smaller intervals, and (2) the amount of phase change over the interval  $T_p$ , which was naturally limited to  $|\alpha_n| \leq \pi$  in Eq. (1.21), is now allowed to exceed this limit as long as the average spectral content satisfies a defined spectral mask (e.g., a Gaussian PSD via Eq. 1.14). Again setting  $BT = 64$ , Fig. 1.16 illustrates the autocorrelation for an optimized overcoded waveform in which the PSL value is  $-52.0$  dB; an improvement of almost  $13$  dB over the LPM bound.

Generally speaking, there is a continuum of possible continuous phase functions, and thus in theory there are an infinite number of possible waveforms for a finite  $BT$ . Besides the PCFM and variants discussed before, there has also been recent work on the use of Bézier curves [37], the Zak transform [27], and polynomial functions [28], as well as various piecewise NLFM structures. Notwithstanding quantization and system error effects, and for a given  $BT$ , it remains to be seen just how low the autocorrelation sidelobes can be made for FM waveforms.

**FIG. 1.16**

Autocorrelation of an optimized overcoded PCFM waveform for  $BT = 64$  [97].

### 1.4.2 SPECTRUM-SHAPED FM WAVEFORMS

It is well known [21] that waveforms for which the PSD decreases toward the band edges achieve low autocorrelation sidelobes. Given the Fourier relationship between autocorrelation and PSD, the waveform design problem can thus be posed with respect to some desired PSD, such as via the frequency domain metric defined in Eq. (1.14). Here we consider an alternating projection approach to perform this PSD-based design that enables the practical realization of ultralow sidelobe (ULS) waveforms [34].

The method of alternating projections has been well studied in the literature (e.g., see Ref. [98] and references therein). A particular case of interest to waveform design is the Gerchberg-Saxton (GS) algorithm [99], variations of which have been used to design unimodular sequences [100–103] and MIMO codes [104]. Here we use a GS-type algorithm to design an FM waveform jointly with a low-loss amplitude taper. The focus on FM structure is to ensure the waveform is amenable to a physical radar transmitter by avoiding phase discontinuities (per Section 1.3.1). The jointly designed amplitude taper then necessitates some modification to the transmitter as will be discussed shortly. This capability is demonstrated with both simulated and experimental measurements.

The first priority for this design approach [34] is to select a PSD  $|G(f)|^2$  defined over some frequency interval  $[f_L, f_H]$ , whose associated autocorrelation possesses low range sidelobes. The defined frequency interval should capture the desired 3-dB bandwidth as well as a sufficient portion of the spectral roll-off such that, when discretized, the adjacent samples avoid abrupt phase changes. Noting that a signal with finite time support cannot actually be bandlimited, this “oversampling” is needed to approximate the continuous FM waveform in a sampled form with sufficient fidelity. The Gaussian PSD depicted in Fig. 1.7 is useful because it provides good spectral containment for this “oversampling” and the associated inverse Fourier transform likewise yields a Gaussian-shaped autocorrelation. It is also necessary to select an initial amplitude taper that is consistent with the desired spectral roll-off, which for the most part is determined by the pulse rise/fall-time [33]. A Tukey taper is a sufficient initialization that is convenient because it permits selection of the degree of amplitude roll-off.

This waveform optimization process is performed in two stages. The first stage iteratively applies the GS algorithm using the selected spectral shape and amplitude taper, where the latter also serves as a soft constraint to minimize SNR loss by maintaining the amplitude envelope close to constant (aside from the tapered roll-off at the pulse edges). The second stage then iterates on the spectral shape alone to fine-tune the match to the PSD, and thus the desired autocorrelation. In so doing, the amplitude envelope can deviate to a small degree from the initial amplitude taper, with this additional design freedom permitting significant further sidelobe reduction with negligible impact to SNR loss (since the optimization will already have converged quite close to the desired PSD by this point).

The first stage involves the alternating application of [34]

$$r_{i+1}(t) = \mathbb{F}^{-1}\{|G(f)| \exp(j\angle \mathbb{F}\{p_i(t)\})\} \quad (1.28)$$

and

$$p_{i+1}(t) = w(t) \exp(j\angle r_{i+1}(t)), \quad (1.29)$$

where  $p_0(t)$  is the initializing FM waveform,  $w(t)$  is the desired amplitude taper,  $\mathbb{F}\{\bullet\}$  is the Fourier transform,  $\mathbb{F}^{-1}\{\bullet\}$  is the inverse Fourier transform, and  $\angle(\bullet)$  extracts the phase of the argument. These steps are repeated  $I$  times to produce the output waveform for the first stage, denoted  $p_I(t)$ , which has both FM and AM attributes. Based on the conservation of ambiguity in Section 1.2, with regard to low range sidelobes being enabled by chirp-like waveforms, it stands to reason that the initial waveform  $p_0(t)$  should have chirp-like qualities to aid convergence. Further, it is convenient to make the 3-dB bandwidth of this initial waveform wider than that of the desired PSD that is providing the spectral shaping.

For the second stage, define the initial waveform as  $q_0(t) = p_I(t)$  using the final result from the previous stage. Then repeat [34]

$$q_{k+1}(t) = x(t) \times \mathbb{F}^{-1}\{|G(f)| \exp(j\angle \mathbb{F}\{q_k(t)\})\} \quad (1.30)$$

$K$  times to produce the final ULS waveform  $s(t) = q_K(t)$ , where  $x(t)$  is a rectangular window of time support  $T$  that serves to limit the temporal extent of  $q_{k+1}(t)$ . Note that when discretized the multiplication  $\times$  becomes a Hadamard product.

For example, consider the design of a waveform having 3-dB bandwidth  $B = 55$  MHz and pulsedwidth  $T = 1.6$   $\mu$ s, such that  $BT = 88$ . The PSD is selected to have a Gaussian shape with the same 3-dB bandwidth and a Tukey taper with amplitude roll-off in the first and last 50 ns is used to shape the pulse rise/fall-time in the first stage. The discretized waveform representation is sampled at about 3.6 times the desired 3-dB bandwidth. The waveform is optimized for  $I = K = 5000$  iterations in each of the two stages. The initial waveform  $p_0(t)$  is derived from the chirp-like ( $L = 8, M = 2$ ) result from Ref. [97] having  $BT = 64$  and PSL =  $-52$  dB, that here has been interpolated up to  $B = 80$  MHz (or  $BT = 128$ ) using polynomial fitting.

Fig. 1.17 depicts the amplitude envelope (in dB scale) of the intermediate waveform (after first stage optimization) and the final ULS waveform (after second stage optimization), along with the Tukey taper used in the first stage. While the intermediate waveform still matches the Tukey taper (they are indistinguishable), the final ULS waveform possess some small additional AM effects that collectively yield an SNR loss (relative to an untapered pulse) of only 0.26 dB.

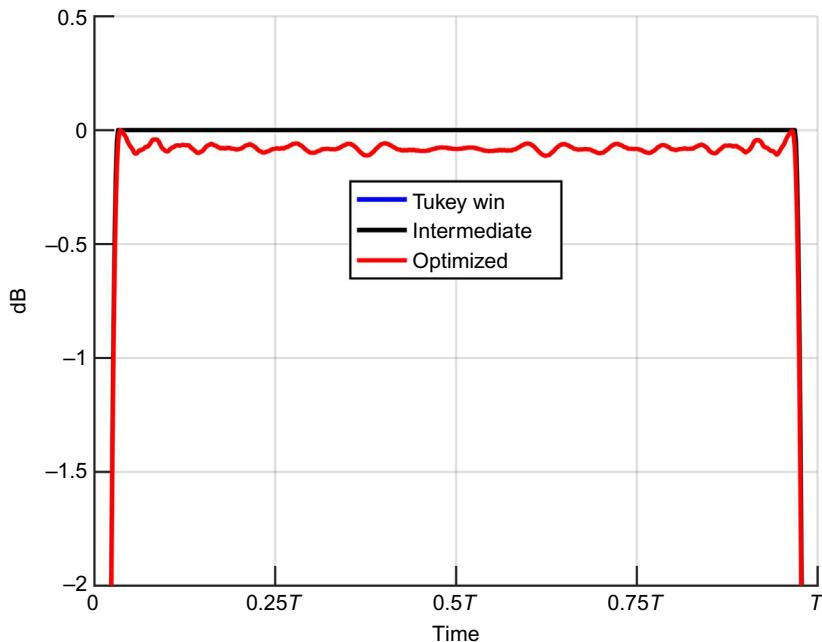


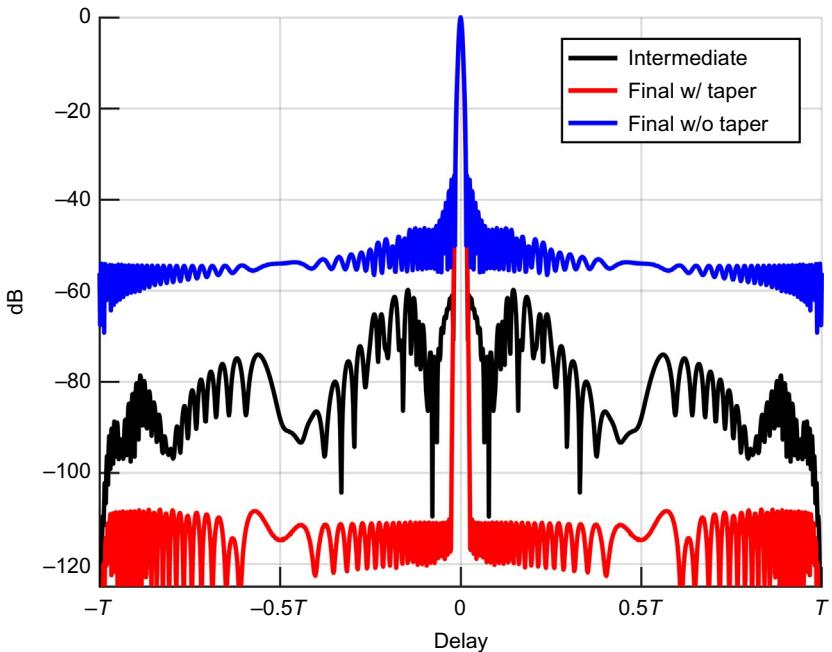
FIG. 1.17

Amplitude envelope of optimized waveform with  $BT = 88$  [34].

[Fig. 1.18](#) shows the autocorrelations for the intermediate and ULS waveforms, along with the autocorrelation of the ULS waveform with the jointly designed amplitude tapering removed. Where the intermediate stage realizes a PSL of  $-59.9$  dB, the ULS waveform achieves an astounding  $-108.1$  dB. However, more than 60 dB of this sidelobe reduction capability is lost if the amplitude tapering is removed.

[Fig. 1.19](#) illustrates the spectral content of the three versions of waveforms from [Fig. 1.18](#). Not only does the low-loss amplitude taper facilitate greatly reduced range sidelobes, it also is clearly necessary to achieve the specified containment in spectral roll-off. Lastly, the ULS waveform also retains its chirp-like structure, as can be observed from the range-Doppler ambiguity function in [Fig. 1.20](#), thus preserving Doppler tolerance.

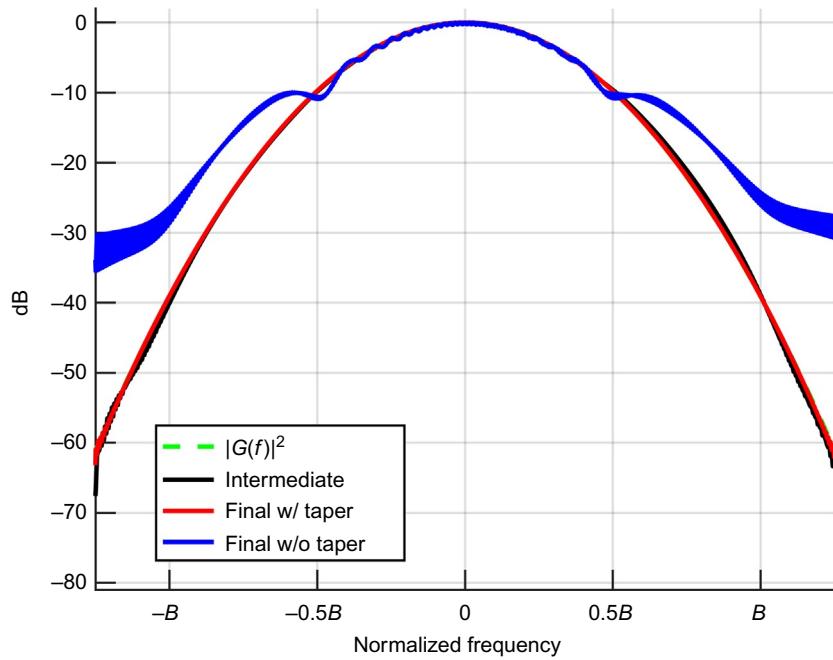
This type of waveform falls within the class of hybrid FM (see also Refs. [35–37]) in which a NLFM waveform is combined with an amplitude taper. The key distinction here is that the FM and AM components of this continuous waveform are optimized jointly, the result of which is tremendously low range sidelobes ([Fig. 1.18](#)), low SNR loss ([Fig. 1.17](#)), good spectral containment ([Fig. 1.19](#)), and Doppler tolerance ([Fig. 1.20](#)).



**FIG. 1.18**

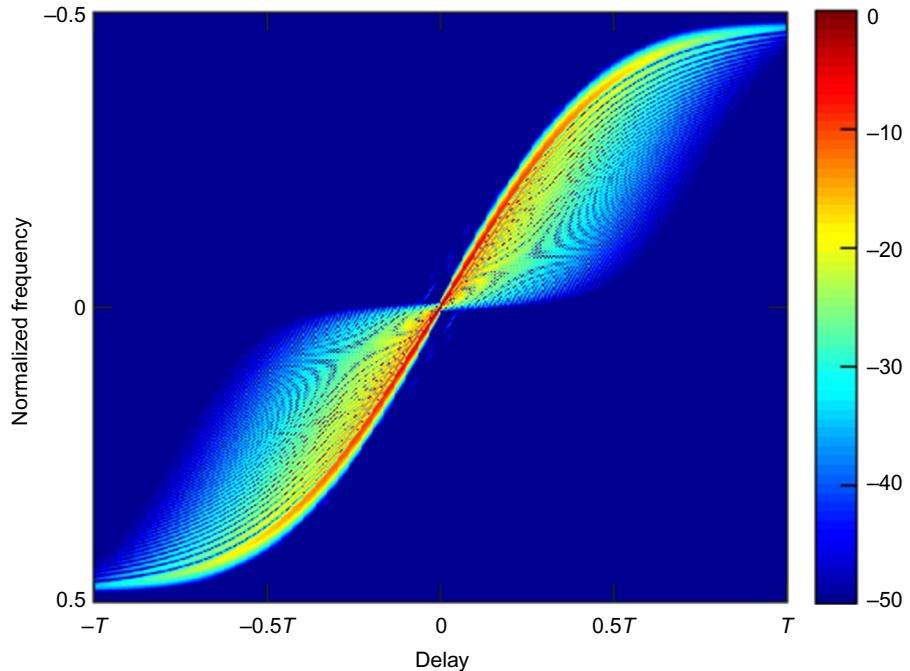
---

Waveform autocorrelation with and without jointly optimized amplitude taper and intermediate waveform autocorrelation [\[34\]](#).



**FIG. 1.19**

Spectral content of ULS waveform with and without jointly optimized amplitude taper and intermediate waveform autocorrelation [34].



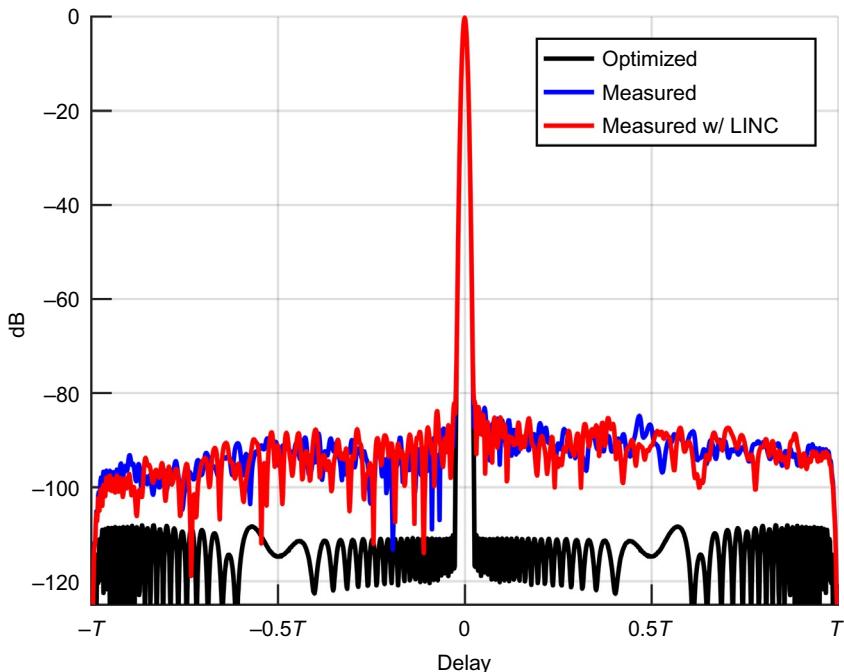
**FIG. 1.20**

Range-Doppler ambiguity function of ULS waveform (in dB) [34].

Of course, the main challenge to realizing all of these benefits is the faithful emission of the waveform given transmitter distortion effects (particularly due to the amplitude saturating HPA) and quantization effects in the waveform generator (presuming it to be digital). While the latter can be addressed by using sufficient bits in the waveform generator, the former requires either operating the HPA in a linear mode with lower power efficiency (may not be possible for some HPAs such as klystrons, TWTs, etc.) or using some manner of predistortion/linearization [31,32].

Upsampling the discretized version of the waveform to 6.125 Gigasamples/s and upconverting to 1.8425 GHz on a Tektronix AWG70002 AWG, a loopback measurement is captured (complex I and Q) by a Rohde & Schwarz spectrum analyzer at 200 Megasamples/s. Measurements were made by (1) directly generating the waveform with the AWG and (2) implemented the waveform using a 180 degrees hybrid coupler via linear amplification using nonlinear components (LINC) [34]. The latter permits the use of two HPAs operated in saturation as long as adequate cross-calibration can be achieved.

[Fig. 1.21](#) shows the autocorrelation for loopback measurements of this waveform captured by the spectrum analyzer (after compensation of linear artifacts [34]). The measurement involving direct use of the AWG attains a PSL of  $-83.2$  dB.



**FIG. 1.21**

Autocorrelations of the direct AWG and LINC-implemented measured ULS waveforms, relative to the ideal optimized version [34].

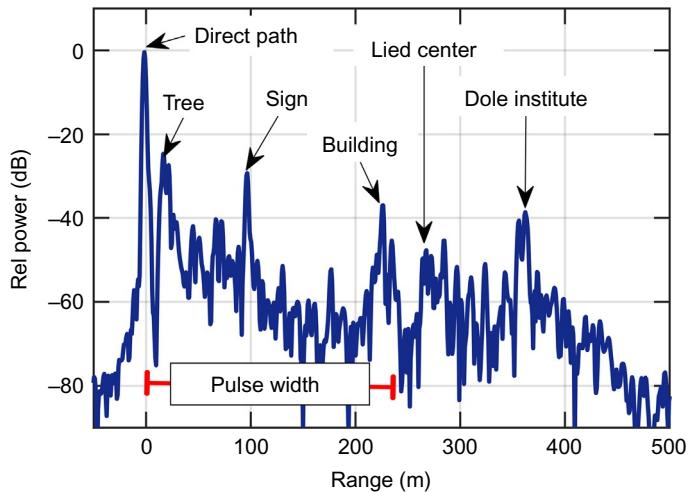
When using the LINC implementation a PSL of  $-81.4$  dB is realized. While still achieving very low sidelobes, the roughly 26-dB degradation relative to the  $-108.1$  dB ideal is largely due to the 10-bit resolution limit of the AWG.

[Fig. 1.22](#) depicts the quasimonostatic test bed (left) and field of view (right) used to made free space measurements with this waveform. By using separate transmit and receive antennas the direct path signal is the dominant “echo” captured by the receiver, as shown in [Fig. 1.23](#). However, the free space emission of the ULS waveform still achieves an apparent dynamic range of around 70 dB. Thus as radar



**FIG. 1.22**

Quasimonostatic test bed (left) and annotated field of view (right) used for rooftop measurements [34].



**FIG. 1.23**

Annotated range profile using ULS waveform [34].

hardware fidelity continues to improve, new enhanced sensitivity waveforms such as ULS can help to facilitate greater detection/discrimination capability.

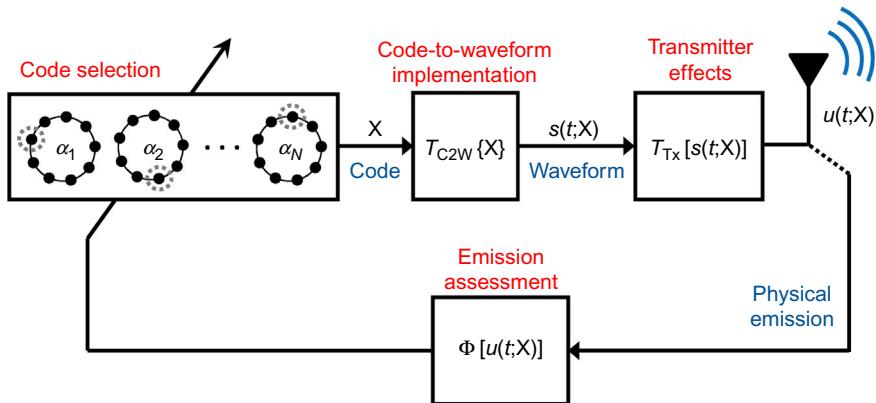
### 1.4.3 TRANSMITTER-IN-THE-LOOP OPTIMIZATION

While FM waveforms and DPSK/MSK implemented binary codes are inherently suitable to radar transmitters [44–47], there is still transmitter-induced distortion that limits sensitivity (per Section 1.3.1). It therefore becomes necessary to consider the impact of the transmitter on the generation of the emitted waveform. To do so, it is useful to employ a stratified nomenclature whereby the *code* (if one exists) encompasses a set of discrete parameters that, when used within a specified implementation scheme (such as DPSK, MSK, or PCFM [47]), produces a continuous *waveform*. This waveform is injected into the transmitter for amplification and subsequent realization of the physical *emission* that is launched from the antenna into the sensing environment.

The most significant source of transmitter distortion is the nonlinearity introduced by the HPA. To compensate for this nonlinearity, many different linearizing transmit architectures have been developed such as the Kahn technique, envelope tracking, various outphasing methods, the Doherty technique, etc. [31]. In contrast, predistortion techniques [32] compensate for nonlinearity by estimating the parameters of models such as Volterra, polynomial, Wiener, and Hammerstein for subsequent inversion of the distortion effects.

Recently, an alternative approach was proposed in Ref. [29] whereby the transmitter distortion is incorporated into the waveform design process. This framework involves two operations: (1) an arbitrary C2W implementation  $s(t; \mathbf{x}) = T_{C2W}\{\mathbf{x}\}$  for code  $\mathbf{x}$ , and (2) the waveform-to-emission transformation  $u(t; \mathbf{x}) = T_{Tx}[s(t; \mathbf{x})]$  imposed by the transmitter distortion. Denote  $\Phi[u(t; \mathbf{x})]$  as the evaluation of the resulting emission using a metric such as PSL, ISL, FTE, etc., described in Section 1.2. Then a “transmitter-in-the-loop” design paradigm (Fig. 1.24) can be used to optimize the radar emission launched into free space. The transmitter distortion transformation can be performed either through a *Model-in-the-Loop* (MiLo) framework employing a mathematical model or through a *Hardware-in-the-Loop* framework that directly uses the radar system under consideration. While the former permit much faster convergence, the latter provides greater accuracy.

Note that the linearization [31] and predistortion [32] architectures could also be used within the transmitter-in-the-loop paradigm, perhaps as a means to achieve joint transmitter/waveform optimization [16] ultimately to mimic the advanced capabilities observed in nature [8]. Such investigation is also necessary to improve radar spectral containment as a means to address the ongoing erosion of radar spectrum [12] due to the rapidly growing demand for wireless services. These joint design approaches have already begun to emerge. Examples include circuit design to optimize jointly for power added efficiency (PAE) and linearization capability [106,107], waveform optimization via Fig. 1.24 within an out-phasing architecture [33], and the development of the *Smith Tube* [108,109] whereby the well-known

**FIG. 1.24**

"Transmitter-in-the-loop" optimization of radar emissions [105].

Smith Chart from RF systems engineering is used to incorporate PAE into waveform design.

As an example of transmitter-in-the-loop optimization, consider the optimization of an emission produced by injecting a waveform into a TWT HPA [110]. The waveform  $s(t)$  is first passed through a fourth-order Chebyshev filter having 3-dB bandwidth that is 2.4 times greater than the waveform that is used to model the linear bandlimiting effects of the transmitter. Denoting the subsequent waveform fed into the TWT as  $s_{in}(t)$ , the resulting amplified output signal is [109]

$$s_{out}(t) = s_{in}(t)A[|s_{in}(t)|] \exp\{j\phi[|s_{in}(t)|]\} \quad (1.31)$$

where  $A[r] = \frac{1}{1 + \beta_a r^2}$  and  $\phi[r] = \frac{a_\phi r^2}{1 + \beta_\phi r^2}$  determine the amount of amplitude and phase distortion, respectively. The term  $a_\phi$  dictates the amount of amplitude-to-phase modulation (also referred to as AM-PM conversion), while  $\beta_a = \beta_\phi = 1/A_s^2$  for  $A_s$  the saturating amplitude.

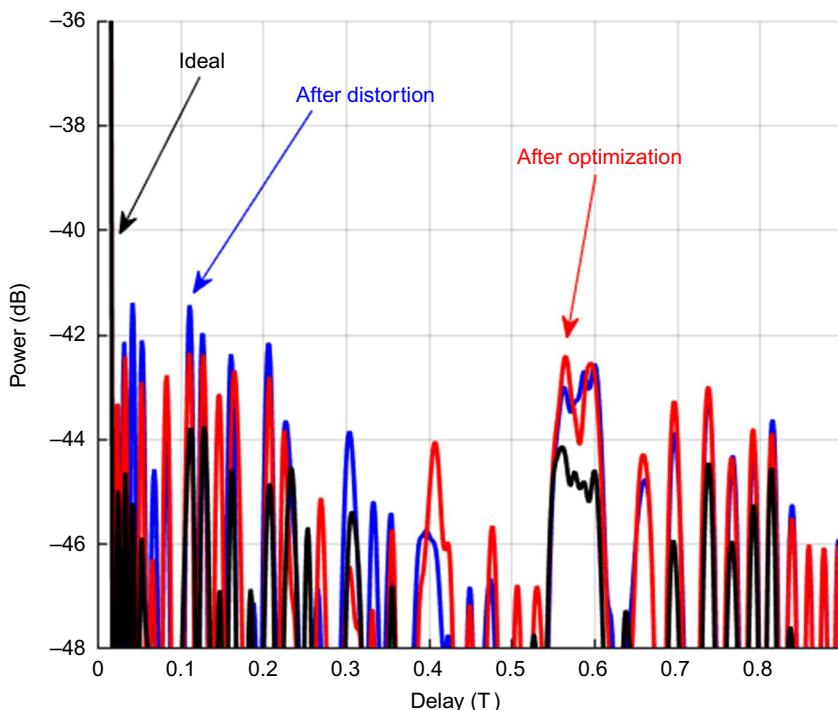
The initial waveform used in this example has a  $BT = 100$  and was obtained using the performance diversity optimization approach of Ref. [29] for an idealistic transmitter (no distortion). This optimized FM waveform has a PSL of  $-43.8$  dB, which is  $0.8$  dB better than the LPM bound from Eq. (1.12). This waveform is injected into the TWT transmitter distortion model for three different operating regimes: mild distortion, moderate distortion, and severe distortion. For each case, MiLo optimization is then performed to compensate for the loss in sensitivity that ensues from the distortion. Per [110], the value  $a_\phi$  is set to  $\pi/12$ , while  $A_s^2$  is set to  $0$ ,  $-5$ , and  $-10$  dB, which are in order of increasing distortion.

**Fig. 1.25** illustrates the mild distortion case ( $A_s^2 = 0$  dB) in which the distorted waveform has a PSL of  $-41.4$  dB, which is a sensitivity loss of  $2.4$  dB. Using the TWT model within the MiLo optimization process yields a PSL of  $-42.4$  dB, a PSL recovery of  $1.0$  dB.

In contrast, **Fig. 1.26** shows the moderate distortion case ( $A_s^2 = -5$  dB) where the PSL is now  $-38.0$  dB, a loss of  $5.8$  dB. For this case MiLo optimization realizes a PSL of  $-41.2$  dB, recovering  $3.2$  dB.

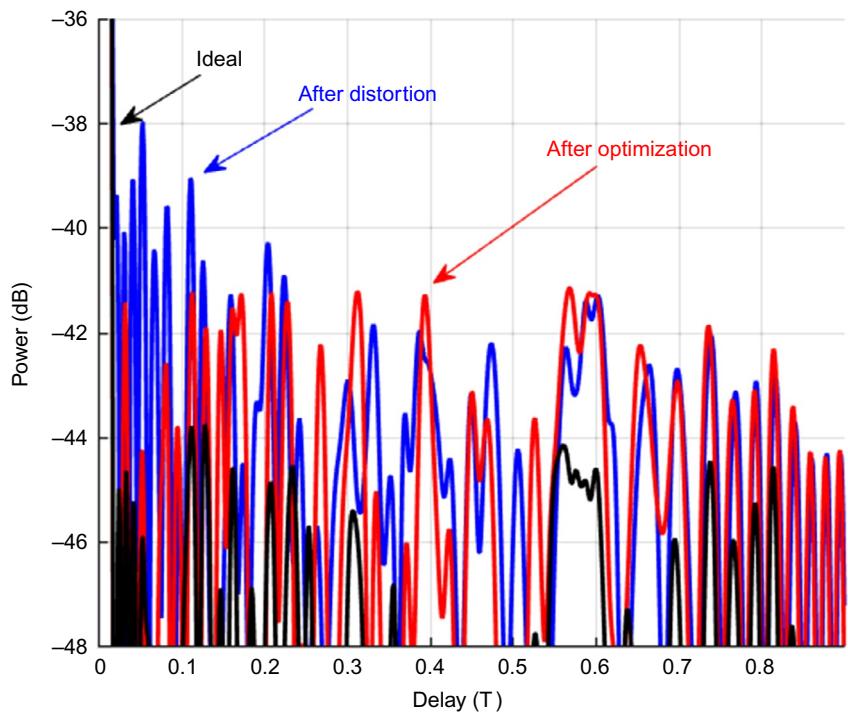
Finally, **Fig. 1.27** depicts the severe distortion case ( $A_s^2 = -10$  dB) where the PSL is now  $-34.2$  dB, a significant loss of  $9.6$  dB. When MiLo optimization is performed using the TWT model, the resulting PSL attained is  $-38.8$  dB, for a recovery of  $4.6$  dB. Clearly, as the distortion becomes more severe, so does the resulting PSL degradation. Performing some manner of transmitter-in-the-loop optimization can compensate for some, but not all, of this sensitivity loss. However, it is expected that the best overall performance will be obtained by some manner of linearization/pre-distortion combined with transmitter-in-the-loop optimization.

Finally, note that the alternating projection waveform design in [Section 1.4.2](#) is not based on an underlying code structure and inherently incorporates a PSD-based



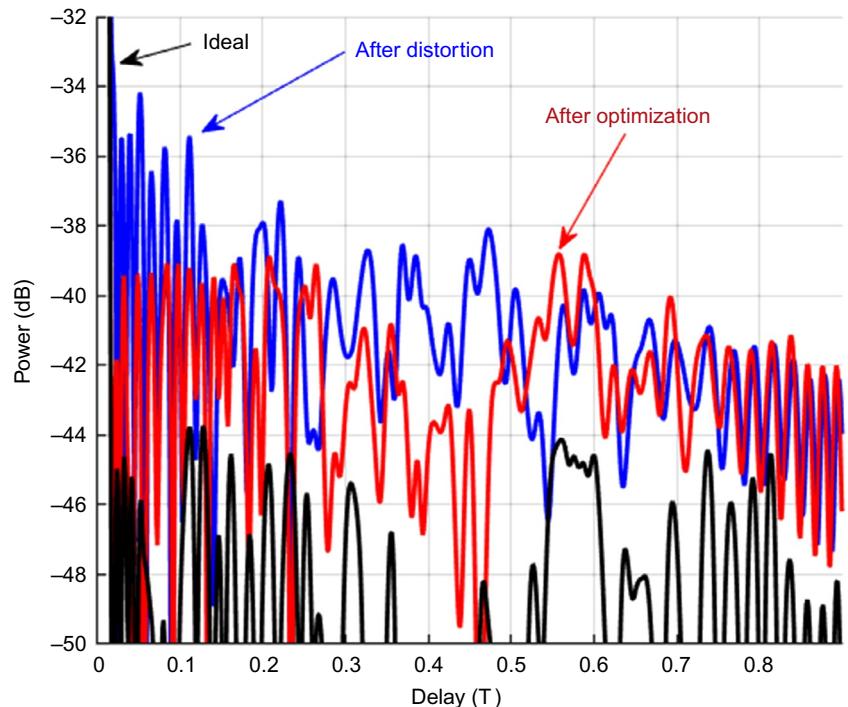
**FIG. 1.25**

Model-in-the-Loop optimization for a TWT HPA (mild distortion) [105].



**FIG. 1.26**

Model-in-the-Loop optimization for a TWT HPA (moderate distortion).



**FIG. 1.27**

Model-in-the-Loop optimization for a TWT HPA (severe distortion) [105].

attribute into the design process. It remains to be seen how this process can be best formulated within the transmitter-in-the-loop optimization structure.

## 1.5 HOLISTIC HIGHER-DIMENSIONAL WAVEFORM DIVERSITY

Leveraging the physically realizable waveform structures discussed in [Section 1.4](#), higher dimensional waveform diversity can now be explored. For example, it was recently shown [\[111\]](#) that the polarization state can be changed in fast-time, thereby realizing joint waveform/polarization modulation. Likewise, new forms of waveform agility have been developed for both CW and pulsed modalities [\[85,88\]](#), along with various associated receive processing schemes [\[82–84,86,87\]](#).

Here we shall focus on the physical realization of spatial domain waveform diversity, otherwise known as MIMO. In so doing, it is important to consider the practical electromagnetic effects of mutual coupling [\[66,67\]](#) and the “emission” of power into what is referred to as the *invisible space* [\[69,70\]](#). Specifically, two particular forms are examined: (1) fast-time spatial modulation that mimics the actuation of the human eye [\[112,113\]](#), and (2) the design of wideband MIMO emissions [\[71,72\]](#).

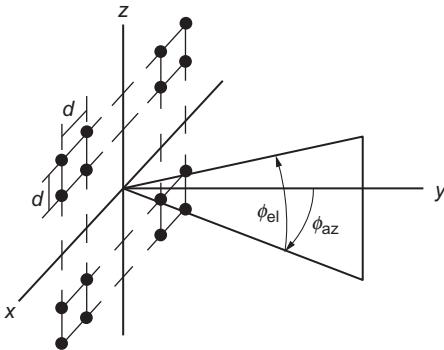
### 1.5.1 SPATIAL MODULATION

The PCFM C2W implementation discussed in [Section 1.4.1](#) can be extended to facilitate a particular form of physically realizable MIMO radar that mimics fixational eye movement in human vision (and in other animals that possess fovea) [\[114,115\]](#). Because this approach maintains a coherently focused beam in a spatial direction that is modulated in fast-time over a small region, it is less susceptible than “orthogonal” waveform MIMO to mutual coupling-induced array pattern errors [\[66,67\]](#), voltage standing wave ratio fluctuations, and near-field energy storage effects [\[69,70\]](#).

Because of the relation to the spatial “jittering” of the eye, which has been linked to visual acuity [\[114,115\]](#), this emission scheme may potentially have application to radar tracking. Further, it has recently been shown [\[71,72\]](#) that the resulting expansion of dimensionality enabled by such delay-angle coupled emissions provides greater degrees of freedom that can be exploited by adaptive receive processing [\[116\]](#). It is also worth noting that this coded FM form of MIMO actually subsumes the related notion of the frequency diverse array [\[117–120\]](#). Here we summarize the 2D spatial configuration that could be generalized to arbitrary planar array structures or alternatively simplified to a linear array for 1D spatial modulation.

Consider the uniform planar array with half-wavelength spacing ( $d = \lambda/2$ ) in [Fig. 1.28](#). The azimuth and elevation angles  $\phi_{az}$  and  $\phi_{el}$ , respectively, are measured relative to array boresight, where they are defined as  $(0^\circ, 0^\circ)$ . With respect to the center of the array, the array elements are indexed as  $(m_x, m_z)$ , in which

$$\begin{aligned} m_x &= -(M_x - 1)/2, -(M_x - 1)/2 + 1, \dots, (M_x - 1)/2 \\ m_z &= -(M_z - 1)/2, -(M_z - 1)/2 + 1, \dots, (M_z - 1)/2 \end{aligned} \quad (1.32)$$

**FIG. 1.28**

Uniform planar array geometry.

for  $M_x$  and  $M_z$  the number of horizontal and vertical array elements, respectively.

First, recall from [Section 1.4.1](#) the length  $N$  phase-change code  $\mathbf{x}_w = [\alpha_1 \ \alpha_2 \ \dots \ \alpha_N]^T$ , here with the subscript “w” added to denote waveform modulation, that is related to a length  $N+1$  polyphase code  $[\theta_0 \ \theta_1 \ \theta_2 \ \dots \ \theta_N]$  via Eqs. [\(1.21\)](#) and [\(1.22\)](#). To perform 2D fast-time spatial modulation, two additional length  $N+1$  codes are needed, which we shall define as  $\Delta_{az,0}, \Delta_{az,1}, \dots, \Delta_{az,N}$ , and  $\Delta_{el,0}, \Delta_{el,1}, \dots, \Delta_{el,N}$  for azimuth and elevation beamsteering, respectively. These code values correspond to spatial angle offsets relative to some center look direction in azimuth ( $\phi_{az,c}$ ) and elevation ( $\phi_{el,c}$ ).

Leveraging the PCFM implementation and assuming that the spatial modulation does not “wrap around” the endfire directions, the spatial phase-change values are obtained as [\[112,113\]](#)

$$\varepsilon_{x,n} = \frac{2\pi d}{\lambda} \sin(\phi_{az,c} + \Delta_{az,n}) \cos(\phi_{el,c} + \Delta_{el,n}) - \frac{2\pi d}{\lambda} \sin(\phi_{az,c} + \Delta_{az,n-1}) \cos(\phi_{el,c} + \Delta_{el,n-1}) \quad (1.33)$$

and

$$\varepsilon_{z,n} = \frac{2\pi d}{\lambda} [\sin(\phi_{el,c} + \Delta_{el,n}) - \sin(\phi_{el,c} + \Delta_{el,n-1})], \quad (1.34)$$

noting that Eqs. [\(1.33\)](#) and [\(1.34\)](#) are expressed in terms of electrical angle (though the term “spatial” will be used here to avoid confusion with the waveform modulation). These values are collected into the length  $N$  spatial phase-change codes denoted  $\mathbf{x}_{s,x} = [\varepsilon_{x,1} \ \varepsilon_{x,2} \ \dots \ \varepsilon_{x,N}]^T$  and  $\mathbf{x}_{s,z} = [\varepsilon_{z,1} \ \varepsilon_{z,2} \ \dots \ \varepsilon_{z,N}]^T$  for the horizontal and vertical array dimensions, respectively. These codes then facilitate the continuous, spatially modulating signals

$$b_x(t; \mathbf{x}_{s,x}) = \exp \left\{ -j \left( \int_0^t g(\tau)^* \left[ \sum_{n=1}^N \varepsilon_{x,n} \delta(\tau - (n-1)T_p) \right] d\tau + \bar{\Delta}_{x,0} \right) \right\} \quad (1.35)$$

and

$$b_z(t; \mathbf{x}_{s,z}) = \exp \left\{ -j \left( \int_0^t g(\tau)^* \left[ \sum_{n=1}^N \epsilon_{z,n} \delta(\tau - (n-1)T_p) \right] d\tau + \bar{\Delta}_{z,0} \right) \right\}, \quad (1.36)$$

where  $\bar{\Delta}_{x,0} = \frac{2\pi d}{\lambda} \sin(\phi_{az,c} + \Delta_{az,0}) \cos(\phi_{el,c} + \Delta_{el,0})$  and  $\bar{\Delta}_{z,0} = \frac{2\pi d}{\lambda} \sin(\phi_{el,c} + \Delta_{el,0})$  are the initial azimuth and elevation offsets (in terms of electrical angle) relative to the center look direction.

Combining the spatially modulating signals of Eqs. (1.35) and (1.36) with the PCFM modulated waveform from Eq. (1.23) parameterized with  $\mathbf{x}_w$  yields a set of  $M_x M_z$  signals driving the  $M_x \times M_z$  elements of the planar array. Transmitter distortion effects notwithstanding, the  $(m_x, m_z)$  array element emits the signal

$$\begin{aligned} s_{m_x, m_z}(t; \mathbf{x}_w, \mathbf{x}_{s,x}, \mathbf{x}_{s,z}) &= s(t; \mathbf{x}_w) b_x^{m_x}(t; \mathbf{x}_{s,x}) b_z^{m_z}(t; \mathbf{x}_{s,z}) \\ &= \exp \left\{ j \left( \int_0^t g(\tau)^* \left[ \sum_{n=1}^N \tilde{\alpha}_n(m_x, m_z) \delta(\tau - (n-1)T_p) \right] d\tau + \tilde{\theta}_0(m_x, m_z) \right) \right\}, \end{aligned} \quad (1.37)$$

where

$$\begin{aligned} \tilde{\alpha}_n(m_x, m_z) &= \alpha_n - m_x \epsilon_{x,n} - m_z \epsilon_{z,n} \\ \tilde{\theta}_0(m_x, m_z) &= \theta_0 - m_x \bar{\Delta}_{x,0} - m_z \bar{\Delta}_{z,0} \end{aligned} \quad (1.38)$$

according to the waveform and azimuth/elevation spatial codes and the array element indices  $m_x$  and  $m_z$ . Note that, despite the complicated coding, the structure in Eq. (1.37) is still just an FM waveform and thus is amenable to a radar transmitter.

As a function of azimuth and elevation, the baseband representation of the far-field emission produced by this  $M_x \times M_z$  array of signals from Eq. (1.37) takes the form

$$g(t, \phi_{az}, \phi_{el}) = \frac{1}{M_x M_z} \sum_{m_x} \sum_{m_z} s_{m_x, m_z}(t) e^{j(k_x(\phi_{az}, \phi_{el}) m_x + k_z(\phi_{el}) m_z)}, \quad (1.39)$$

where  $k_x = 2\pi d \sin(\phi_{az}) / \lambda$  and  $k_z = 2\pi d \sin(\phi_{el}) / \lambda$ , for  $\lambda$  the wavelength of the center frequency, and the code dependencies have been suppressed for brevity. The structure of the spatial modulation, averaged over the pulsedwidth, can be evaluated using the aggregate beampattern defined as

$$\text{aggregate beampattern} = \frac{1}{T} \int_0^T g(t, \phi_{az}, \phi_{el}) g^*(t, \phi_{az}, \phi_{el}) dt. \quad (1.40)$$

As an example, consider a length  $N = 100$  waveform code  $\mathbf{x}_w$  that closely approximates an LFM waveform as described in Eq. (1.24). The planar array has 8 horizontal elements ( $M_x = 8$ ) and 12 vertical elements ( $M_z = 12$ ), with half-wavelength spacing and boresight center direction:  $(\phi_{az,c}, \phi_{el,c}) = (0^\circ, 0^\circ)$ . Fig. 1.29 illustrates the aggregate beampatterns obtained for three different cases. The first (Fig. 1.29A)

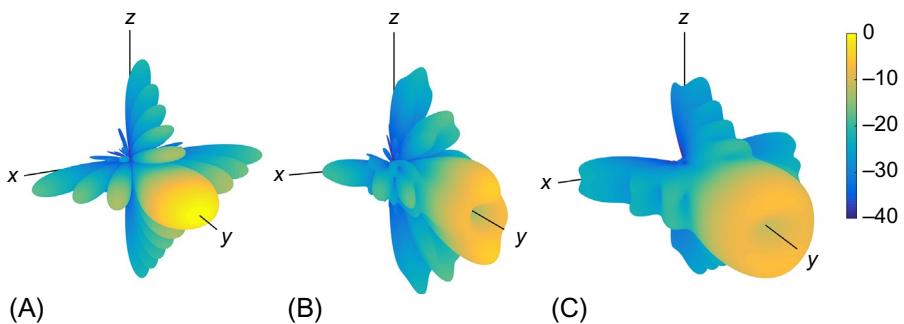


FIG. 1.29

Aggregate beampattern in Cartesian coordinates for (A) standard beamforming; (B) phase-only beam-spoiling; and (C) one revolution of circular spatial modulation [113].

involves standard beamforming in the boresight direction (so no spatial modulation), which serves as a point of reference for comparison. The second case (Fig. 1.29B) employs a phase-only weighting across the array (*not* time varying) to achieve a beam-spoiling effect that widens the beam. Such an approach provides greater flexibility in beam shaping while still permitting operation of the HPAs in saturation to maximize emitted power (e.g., see Refs. [121,122]).

The MIMO structure can be viewed as providing a time-varying form of beam-spoiling which, when taken in aggregate, yields greater degrees of freedom for beam shaping (e.g., see Refs. [123,124]). Fig. 1.29C illustrates just such an example whereby the spatial modulation framework is used to traverse a single circle of spatial-angle radius 9.6 degrees (the first null in elevation angle) during the pulselength. Because it is based on the physically realizable PCFM structure [47], this emission scheme is readily amenable to high power radar and the underlying coded structure (via  $\mathbf{x}_w$ ,  $\mathbf{x}_{s,x}$ , and  $\mathbf{x}_{s,z}$ ) permits optimization via a variety of different approaches.

As expected based on the virtual array concept [61], spatial modulation has been found to provide a spatial resolution enhancement of as much as 30% compared to standard phased array beamforming [112]. Further, certain combinations of waveform/spatial modulation realize what amounts to a far-field tapering effect that reduces spatial sidelobes. However, the trade-off for these improvements is a degradation in range resolution that arises from imparting only portions of the overall bandwidth in each spatial direction as the beam sweeps. As such, the practical applications for such an approach would logically involve those in which the degree of spatial modulation would be kept relatively small (e.g., tracking).

### 1.5.2 HOLISTIC WIDEBAND MIMO RADAR

Finally, we consider the holistic design of wideband MIMO emissions which necessitates taking a joint space-frequency perspective that accounts for spectral content and electromagnetic near-field energy storage to avoid damage to the transmitter that

could arise from large amounts of reactive power [69,70]. The solution here again leverages the alternating projection structure of the GS algorithm [99] discussed in [Section 1.4.2](#), which has previously been applied to MIMO radar [104] as well as wideband beampattern synthesis [125]. Here the GS formulation is used for wideband MIMO radar design to alternate between the element-time and space-frequency domains to obtain transmitter-amenable FM waveforms that minimize reactive power [71,72]. In so doing, this emission design enables the physical realization of wideband MIMO SAR [126] and simultaneous multimode operation [127].

Before addressing wideband MIMO emission design, it is first necessary to establish the nomenclature for wideband beamforming. For a uniform linear array, it is generally desirable for the interelement antenna spacing  $d$  to be a half-wavelength to avoid grating lobes [68]. Defining this particular wavelength as  $\lambda_d$ , we can thus write

$$d = \frac{\lambda_d}{2} = \frac{c}{2f_d} \quad (1.41)$$

where  $c$  is the speed of light and  $f_d$  is the associated frequency. It can therefore be shown that the electrical angle  $\bar{\phi}$  as a function of frequency and spatial angle  $\phi$  can be expressed as

$$\bar{\phi}(f, \phi) = \pi \frac{f}{f_d} \sin \phi, \quad (1.42)$$

which is a convenient way to define electrical angle for wideband operation.

For an  $M$  element uniform linear array, let  $s_m(t)$  be the continuous, pulsed waveform emitted from the  $m$ th antenna element, the discretized version of which is the length  $N$  vector  $\mathbf{s}_m$  that is “oversampled” with respect to some nominal bandwidth  $B$  to capture enough spectral roll-off to ensure sufficient fidelity. Collecting the  $M$  discretized waveforms into the  $N \times M$  matrix  $\mathbf{S}$ , the collective complex-baseband emission in spatial direction  $\phi$  at frequency  $f$  can be determined via

$$g(f, \phi) = \frac{1}{M} \mathbf{a}^H(f) \mathbf{S} \mathbf{v}(f, \phi), \quad (1.43)$$

where

$$\mathbf{v}(f, \phi) = \left[ 1 \quad e^{j\bar{\phi}(f, \phi)} \quad \dots \quad e^{j(M-1)\bar{\phi}(f, \phi)} \right]^T \quad (1.44)$$

is the  $M \times 1$  frequency-dependent steering vector. Likewise,

$$\mathbf{a}(f) = \left[ 1 \quad e^{j2\pi \frac{f-f_{\text{cent}}}{f_{\text{samp}}}} \quad \dots \quad e^{j2\pi(N-1)\frac{f-f_{\text{cent}}}{f_{\text{samp}}}} \right]^T \quad (1.45)$$

is the  $N \times 1$  discrete-time Fourier transform vector as a function of continuous passband frequency  $f$ , for  $f_{\text{samp}}$  the sampling rate of the discretized waveform in  $\mathbf{S}$  and  $f_{\text{cent}}$  the passband center frequency.

Also note that the (unitless) *fractional bandwidth* is

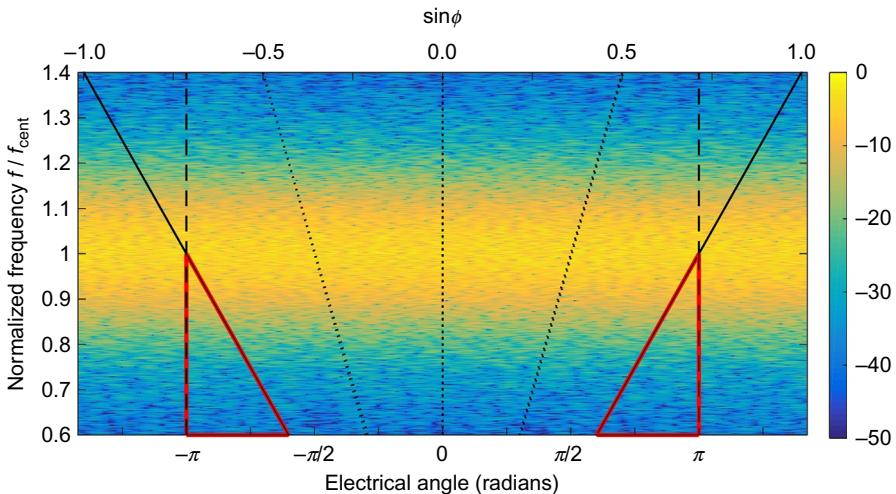
$$\% \text{BW} = \frac{B}{f_{\text{cent}}}. \quad (1.46)$$

To provide control over out-of-band power, in this context we define bandwidth  $B$  in terms of the frequency interval comprising 98% of the power, which is a *percent power bandwidth* in units of Hertz not to be confused with Eq. (1.46). A fractional bandwidth greater than 10% is generally considered to be wideband [128,129].

For narrowband operation, we have  $f \approx f_d$  so that Eq. (1.31) simplifies to  $\bar{\phi}(f, \phi) \cong \pi \sin \phi$ , where the spatial angles at endfire ( $\phi = \pm \pi/2$ ) map to the electrical angles  $\bar{\phi} = \pm \pi$ . However, when  $f < f_d$  (corresponding to  $\lambda > \lambda_d$ ) each endfire spatial angle coincides with an electrical angle for which  $|\bar{\phi}| < \pi$ . The electrical angles beyond this point do not “wrap around” the endfire direction but instead correspond to the *imaginary space* (or *invisible space*) [69] associated with reactive power and near-field energy storage. This power can be reflected back into the transmitter, subsequently damaging the radar [70]. Fig. 1.30 illustrates the example of a MIMO emission with 30% fractional bandwidth, where the *lower triangles* demarcate the invisible space regions. We wish to design MIMO waveforms such that the resulting emission imparts minimal power to these regions.

A GS-based alternating projection approach can be applied to minimize the power in the invisible space by designing the MIMO waveforms to realize a far-field joint space-frequency emission that maximizes the power in the real space. For an “oversampling” factor of  $\kappa$  relative to 3-dB bandwidth (such as discussed in Section 1.4.2), we have  $f_{\text{samp}} = \kappa B$  and the baseband frequency domain is discretized into  $Q$  equally spaced points according to

$$f_q = \frac{-f_{\text{samp}}}{2} + \frac{q}{Q} f_{\text{samp}} \quad (1.47)$$



**FIG. 1.30**

Wideband frequency content versus electrical angle for typical MIMO emission. Element spacing is half-wavelength for center frequency  $f_{\text{cent}}$  (so  $d = 0.5\lambda_{\text{cent}}$ ) with 30% fractional bandwidth. *Lower triangles* identify invisible space and diagonal lines indicate  $\sin(\phi)$  [71].

for  $q=0, 1, \dots, Q-1$ . The discretized passband frequencies are thus  $f=f_q+f_{\text{cent}}$ . Likewise,  $\phi_p$  for  $p=0, 1, \dots, P-1$  are the spatial angles denoted as “beamlets” that are to be included in the optimization, which could include a discretization of all spatial angles from endfire to endfire or some subset thereof depending on the application. This space-frequency discretization thus leads to the discretized version of Eq. (1.43) as

$$g_i(f_q, \phi_p) = \frac{1}{M} \mathbf{a}^H(f_q) \mathbf{S}_{i-1} \mathbf{v}(f_q, \phi_p), \quad (1.48)$$

where we have also introduced the dependence of  $g_i(f_q, \phi_p)$  in the  $i$ th iteration on the previous discretized waveform matrix  $\mathbf{S}_{i-1}$ . From Eq. (1.48) we can also construct the  $Q \times P$  space-frequency matrix

$$\mathbf{G}_i = \begin{bmatrix} g_i(f_0, \phi_0) & g_i(f_0, \phi_1) & \dots & g_i(f_0, \phi_{P-1}) \\ g_i(f_1, \phi_0) & g_i(f_1, \phi_1) & \dots & g_i(f_1, \phi_{P-1}) \\ \vdots & \vdots & \ddots & \vdots \\ g_i(f_{Q-1}, \phi_0) & g_i(f_{Q-1}, \phi_1) & \dots & g_i(f_{Q-1}, \phi_{P-1}) \end{bmatrix}. \quad (1.49)$$

Now define the (magnitude) spectral window for the  $p$ th beamlet as the  $Q \times 1$  vector  $\mathbf{u}(\phi_p) = [u(f_0, \phi_p) \ u(f_1, \phi_p) \ \dots \ u(f_{Q-1}, \phi_p)]^T$ , the purpose of which is to shape the spectrum according to a desired PSD, inclusive of the spectral roll-off as discussed in Section 1.4.2. An adaptive scaling denoted as  $b(\phi_p)$  is utilized as  $b(\phi_p) \mathbf{u}(\phi_p)$  to provide the optimization process with sufficient freedom to control the desired power distribution as a function of spatial angle. The collection of  $Q \times P$  spectral shaping vectors can thus defined as

$$\mathbf{U} = [b(\phi_0) \mathbf{u}(\phi_0) \ b(\phi_1) \mathbf{u}(\phi_1) \ \dots \ b(\phi_{P-1}) \mathbf{u}(\phi_{P-1})], \quad (1.50)$$

with the  $i$ th update of the adaptive scaling determined as

$$b_i(\phi_p) = b_{i-1}(\phi_p) \sqrt{\frac{\mathbf{u}^H(\phi_p) \mathbf{u}(\phi_p)}{\sum_{q=0}^{Q-1} |g_i(f_q, \phi_p)|^2 \sum_{\ell=0}^{P-1} |b_{i-1}(\phi_\ell)|^2}}. \quad (1.51)$$

The spectral shaping is then applied to  $g_i(f_q, \phi_p)$  as

$$\tilde{g}_i(f_q, \phi_p) = b_i(\phi_p) u(f_q, \phi_p) \exp(j\angle g_i(f_q, \phi_p)), \quad (1.52)$$

the  $Q \times P$  collection of which is denoted  $\tilde{\mathbf{G}}_i$ .

Let

$$\mathbf{V}(f_q) = [\mathbf{v}(f_q, \phi_0) \ \mathbf{v}(f_q, \phi_1) \ \dots \ \mathbf{v}(f_q, \phi_{P-1})] \quad (1.53)$$

be the  $M \times P$  steering vector matrix for frequency  $f_q$ . The spectrally shaped response is then projected back onto the array to obtain the  $1 \times M$  vector

$$\tilde{\mathbf{r}}_i(f_q) = \tilde{\mathbf{g}}_i(f_q) \mathbf{V}^H(f_q), \quad (1.54)$$

where the  $1 \times P$  vector  $\tilde{\mathbf{g}}_i(f_q)$  is the  $q$ th row of  $\tilde{\mathbf{G}}_i$ . The new unconstrained (in amplitude) discretized waveform matrix can subsequently be found by stacking the  $Q$

frequency dependent  $1 \times M$  vectors from Eq. (1.54) and performing the inverse Fourier transform as

$$\tilde{\mathbf{S}}_i = [\mathbf{a}(f_0) \ \mathbf{a}(f_1) \ \dots \ \mathbf{a}(f_{Q-1})] \begin{bmatrix} \tilde{\mathbf{r}}_i(f_0) \\ \tilde{\mathbf{r}}_i(f_1) \\ \vdots \\ \tilde{\mathbf{r}}_i(f_{Q-1}) \end{bmatrix}. \quad (1.55)$$

Finally, the  $N \times M$  matrix in Eq. (1.55) is forced to be constant amplitude via

$$\mathbf{S}_i = \exp(j\angle\tilde{\mathbf{S}}_i), \quad (1.56)$$

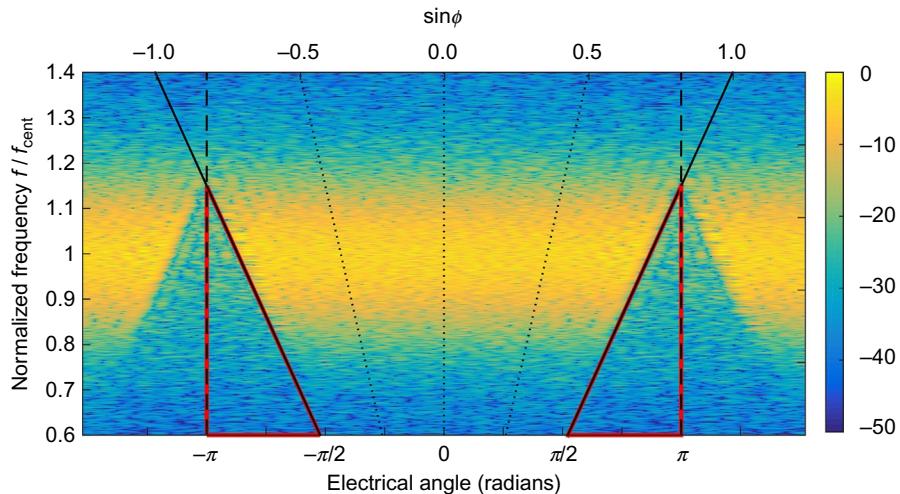
which is understood to be applied on an element-wise basis.

Initializing with some discretized set of waveforms  $\mathbf{S}_0$ , adaptive scaling coefficients  $b_0(\phi_0) = \dots = b_0(\phi_{P-1}) = 1$ , and desired spectral windows  $\mathbf{u}(\phi_0), \mathbf{u}(\phi_1), \dots, \mathbf{u}(\phi_{P-1})$ , the algorithm operates by iteratively applying Eqs. (1.48), (1.51), (1.52), (1.54)–(1.56) until some prescribed convergence criteria is met. In Ref. [72], the incorporation of space-frequency nulling and mapping of the discretized form of Eq. (1.56) into the PCFM waveform structure are also discussed.

As an example, consider a near-omnidirectional MIMO emission using  $M = 30$  elements, waveforms with  $BT = 100$ , and setting  $f_d/f_{\text{cent}} = 1.15$ . Using  $P = 2M = 60$  beamlets distributed as

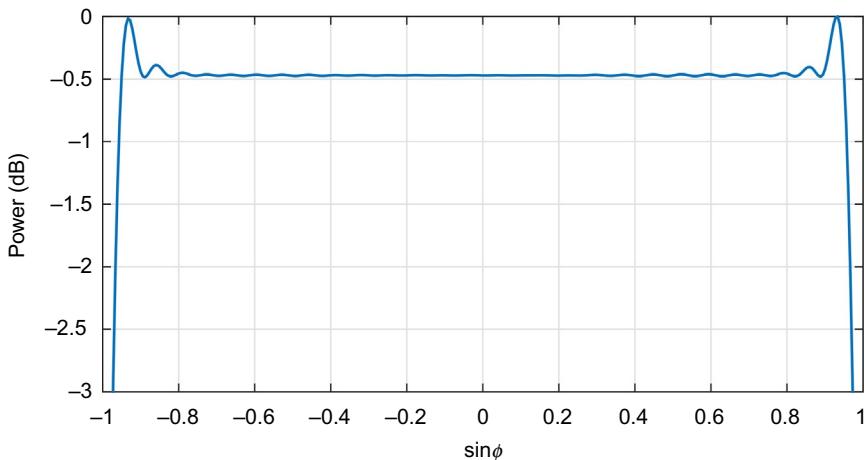
$$\phi_p = \sin^{-1}\left(-1 + 2\frac{(p+0.5)}{P}\right) \quad (1.57)$$

for  $p = 0, 1, \dots, P - 1$ , identical spectral windows having a Gaussian shape, and 30% fractional bandwidth, Fig. 1.31 illustrates the space-frequency response of the

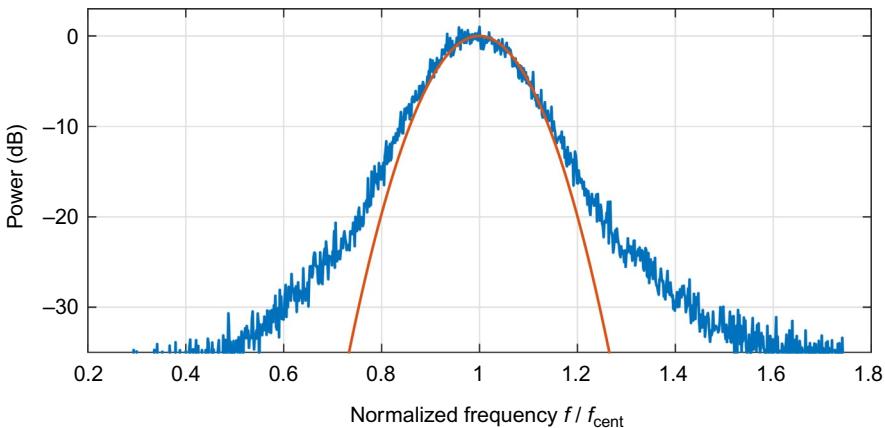


**FIG. 1.31**

Space-frequency response of near-omnidirectional MIMO emission (30% bandwidth) versus electrical angle optimized to minimize reactive power; *diagonal lines* indicate  $\sin(\phi)$  values [71].

**FIG. 1.32**

Peak-normalized near-omnidirectional optimized MIMO beampattern versus  $\sin(\phi)$  [71].

**FIG. 1.33**

For each normalized frequency, maximum spectral power over space  $\phi$  for optimized emission (the broader trace) compared to the PSD template Gaussian spectral window [71].

resulting optimized emission. Since the design process focuses on the far-field emission and, by placing beamlets only in directions that coincide with real space (the selection in Eq. 1.57 also avoids getting too close to the endfire directions), the power “emitted” into the invisible space is clearly quite small.

The distribution of power as a function of spatial angle is depicted in Fig. 1.32, where it is noted that the endfire directions receive little power. Further, aside from a small ripple at the edges, the (peak) power is rather flat across space for this near-omnidirectional emission. Finally, Fig. 1.33 shows the maximum power spectra

taken over spatial angle compared to the desired Gaussian PSD. The desired PSD shape is fit fairly well for higher power but spreads out at the band edges due to the constraint of having constant amplitude waveforms.

---

## 1.6 CONCLUSIONS

To realize waveform diversity in practical radar systems will require, in many cases, the joint consideration of signal processing, systems engineering, and electromagnetics to capture the necessary physical characteristics. Doing so also opens the door to new possibilities for multidimensional optimization and holistic design problems that marry the signal processing to myriad different hardware configurations. This chapter has highlighted some of the recent work taking this holistic perspective for radar. Specifically, approaches to optimize physically realizable FM waveforms have been presented, along with a “transmitter-in-the-loop” design paradigm that facilitates the optimization of the far-field radar emission inclusive of transmitter distortion. Physically realizable forms of MIMO radar have also been discussed that mimic the actuation of the human eye or address the reactive power problem for wideband operation. Given the breadth of signal processing tools juxtaposed against the practical system-level effects, continued spectral issues, and multimode demands on radar, one can expect continued advances into the future.

---

## REFERENCES

- [1] M. Vespe, G. Jones, C.J. Baker, Lesson for radar: waveform diversity in echolocating mammals, *IEEE Signal Process. Mag.* 26 (1) (2009) 65–75.
- [2] T.G. Leighton, S.D. Meers, P.R. White, Propagation through nonlinear time-dependent bubble clouds and the estimation of bubble populations from measured acoustic characteristics, *Proc. R. Soc. Lond. A* 460 (2004) 2521–2550.
- [3] C.J. Baker, G.E. Smith, A. Balleri, M. Holderied, H.D. Griffiths, Biomimetic echolocation with application to radar and sonar sensing, *Proc. IEEE* 102 (4) (2014) 447–458.
- [4] H.U. Schnitzler, E.K.V. Kalko, Echolocation by insect-eating bats, *Bioscience* 51 (2001) 557–569.
- [5] K. Ghose, C.F. Moss, Sonar beam pattern of a flying bat as it tracks tethered insects, *J. Acoust. Soc. Am.* 114 (2) (2003) 1120–1131.
- [6] T.G. Leighton, G.H. Chua, P.R. White, Do dolphins benefit from nonlinear mathematics when processing their sonar returns? *Proc. Roy. Soc. A* 468 (2012) 3517–3532.
- [7] G.H. Chua, P.R. White, T.G. Leighton, Use of clicks resembling those of the Atlantic bottlenose dolphin (*Tursiops truncatus*) to improve target discrimination in bubbly water with biased pulse summation sonar, *IET Radar Sonar Navig.* 6 (6) (2012) 510–515.
- [8] W.W.L. Au, S.W. Martin, Why dolphin biosonar performs so well in spite of mediocre ‘equipment’, *IET Radar Sonar Navig.* 6 (6) (2012) 566–575.
- [9] T.G. Leighton, *The Acoustic Bubble*, Academic Press, London, UK, 1994.

- [10] F.H. Sanders, R.L. Sole, B.L. Bedford, D. Franc, and T. Pawlowitz, Effects of RF Interference on Radar Receivers, NTIA Technical Report, TR-06-444, 2006.
- [11] M.J. Marcus, Spectrum policy for radio spectrum access, Proc. IEEE 100 (2012) 1685–1691.
- [12] H. Griffiths, L. Cohen, S. Watts, E. Mokole, C. Baker, M. Wicks, S. Blunt, Radar spectrum engineering and management: technical and regulatory issues, Proc. IEEE 103 (1) (2015) 85–102.
- [13] M. Wicks, E. Mokole, S.D. Blunt, V. Amuso, R. Schneible (eds.), Principles of Waveform Diversity & Design, SciTech Publishing, Raleigh, NC, USA, 2010.
- [14] S. Pillai, K.Y. Li, I. Selesnick, B. Himed, Waveform Diversity: Theory & Applications, McGraw-Hill, 2011.
- [15] F. Gini, A. De Maio, L.K. Patton, Waveform Design and Diversity for Advanced Radar Systems, IET Press, London, UK, 2012.
- [16] H. Griffiths, S. Blunt, L. Cohen, L. Savy, Challenge problems in spectrum engineering and waveform diversity, in: IEEE Radar Conf., Ottawa, Canada, April/May 2013.
- [17] S.D. Blunt, E.L. Mokole, An overview of radar waveform diversity, IEEE AES Syst. Mag. 31(11) (2016) 2–42.
- [18] N. Levanon, E. Mozeson, Radar Signals, Wiley-IEEE Press, New York, NY, 2004.
- [19] J.R. Klauder, A.C. Price, S. Darlington, W.J. Albersheim, The theory and design of chirp radars, Bell Syst. Tech. J. XXXIX(4) (1960) 745–808.
- [20] W.J. Caputi, Stretch: a time-transformation technique, IEEE Trans. Aerosp. Electron. Syst. AES-7 (2) (1971) 269–278.
- [21] A. Johnston, Improvements to a pulse compression radar matched filter, Radio Electron. Eng. 53 (4) (1983) 138–140.
- [22] B. Manz, Advancing TWTs: the traveling wave tube lives on ... and on, AOC J. Electron. Def., 32, 7, 26-30, 2009.
- [23] O. Holt, Technology survey: a sampling of TWTs and MPMs, AOC J. Electron. Def. 39(3) (2016) 41–47.
- [24] P.M. Woodward, Probability and Information Theory with Applications to Radar, Pergamon Press, New York, NY, 1953.
- [25] C.E. Cook, A class of nonlinear FM pulse compression signals, Proc. IEEE 52 (11) (1964) 1369–1371.
- [26] E. Fowle, The design of FM pulse compression signals, IEEE Trans. Inf. Theory 10 (1) (1964) 61–67.
- [27] I. Gladkova, Design of frequency modulated waveforms via the Zak transform, IEEE Trans. Aerosp. Electron. Syst. 40 (1) (2004) 355–359.
- [28] A.W. Doerry, Generating Nonlinear FM Chirp Waveforms for Radar, Sandia Report, SAND2006-5856, 2006.
- [29] S.D. Blunt, J. Jakabosky, M. Cook, J. Stiles, S. Seguin, E.L. Mokole, Polyphase-coded FM (PCFM) radar waveforms, part II.1: optimization, IEEE Trans. Aerosp. Electron. Syst. 50 (3) (2014) 2230–2241.
- [30] P.S. Tan, J. Jakabosky, J.M. Stiles, S.D. Blunt, On higher-order representations of polyphase-coded FM radar waveforms, in: IEEE Intl. Radar Conf., Arlington, VA, May 2015.
- [31] F.H. Raab, P. Asbeck, S. Cripps, P.B. Kenington, Z.B. Popovic, N. Pothecary, J.F. Sevic, N.O. Sokal, Power amplifiers and transmitters for RF and microwave, IEEE Trans. Microwave Theory Tech. 50 (3) (2002) 814–826.

- [32] F.M. Ghannouchi, O. Hammi, Behavioral modeling and predistortion, *IEEE Microwave Mag.* 10 (7) (2009) 52–64.
- [33] L. Ryan, J. Jakabosky, S.D. Blunt, C. Allen, L. Cohen, Optimizing polyphase-coded FM waveforms within a LINC transmit architecture, in: IEEE Radar Conf., Cincinnati, OH, May 2014.
- [34] J. Jakabosky, S.D. Blunt, T. Higgins, Ultra-low sidelobe waveform design via spectral shaping and LINC transmit architecture, in: IEEE Intl. Radar Conf., Arlington, VA, May 2015.
- [35] J.A. Johnston, A.C. Fairhead, Waveform design and Doppler sensitivity analysis for non-linear FM chirp pulses, *IEE Proc. Commun. Radar Signal Process.* 133 (2) (1986) 163–175.
- [36] T. Collins, P. Atkins, Nonlinear frequency modulation chirps for active sonar, *IEE Proc. Radar Sonar Navig.* 146 (6) (1999) 312–316.
- [37] J. Kurdzo, B.L. Cheong, R. Palmer, G. Zhang, Optimized NLFM pulse compression waveforms for high-sensitivity radar observations, in: Intl. Radar Conf., Lille, France, October 2014.
- [38] B.L. Lewis, F.F. Kretschmer, Linear frequency modulation derived polyphase pulse compression codes, *IEEE Trans. Aerosp. Electron. Syst.* 18 (5) (1982) 637–641.
- [39] C.J. Nunn, G.E. Coxson, Polyphase pulse compression codes with optimal peak and integrated sidelobes, *IEEE Trans. Aerosp. Electron. Syst.* 45 (2) (2009) 775–781.
- [40] R.H. Barker, Group synchronizing of binary digital sequences, in: W. Jackson (ed.), *Communication Theory*, Academic Press, London, UK, 1953, pp. 273–287.
- [41] M.N. Cohen, M.R. Fox, J.M. Baden, Minimum peak sidelobe pulse compression codes, in: IEEE Intl. Radar Conf., Arlington, VA, May 1990.
- [42] G. Coxson, J. Russo, Efficient exhaustive search for optimal-peak-sidelobe binary codes, *IEEE Trans. Aerosp. Electron. Syst.* 41 (1) (2005) 302–308.
- [43] C.J. Nunn, G.E. Coxson, Best-known autocorrelation peak sidelobe levels for binary codes of length 71 to 105, *IEEE Trans. Aerosp. Electron. Syst.* 44 (1) (2008) 392–395.
- [44] H.H. Faust, B. Connolly, T.M. Firestone, R.C. Chen, B.H. Cantrell, E. Mokole, A spectrally clean transmitting system for solid-state phased-array radars, in: IEEE Radar Conf., Philadelphia, PA, April 2004.
- [45] J.W. Taylor, H.J. Blinchikoff, Quadriphase code – a radar pulse compression signal with unique characteristics, *IEEE Trans. Aerosp. Electron. Syst.* 24 (2) (1988) 156–170.
- [46] S. Blunt, M. Cook, E. Perrins, J. de Graaf, CPM-based radar waveforms for efficiently bandlimiting a transmitted spectrum, in: IEEE Radar Conf., Pasadena, CA, May 2009.
- [47] S.D. Blunt, M. Cook, J. Jakabosky, J. de Graaf, E. Perrins, Polyphase-coded FM (PCFM) radar waveforms, part I: implementation, *IEEE Trans. Aerosp. Electron. Syst.* 50 (3) (2014) 2218–2229.
- [48] N. Levanon, Multifrequency complementary phase-coded radar signal, *IEE Proc. Radar Sonar Navig.* 147 (6) (2000) 276–284.
- [49] N. Levanon, E. Mozeson, Multicarrier radar signal – pulse train and CW, *IEEE Trans. Aerosp. Electron. Syst.* 38 (2) (2002) 707–720.
- [50] R.F. Tigrek, W.J.A. De Heij, P. Van Genderen, OFDM signals as the radar waveform to solve Doppler ambiguity, *IEEE Trans. Aerosp. Electron. Syst.* 48 (1) (2012) 130–143.
- [51] G. Lelouch, A.K. Mishra, M. Inggs, Stepped OFDM radar technique to resolve range and Doppler simultaneously, *IEEE Trans. Aerosp. Electron. Syst.* 51 (2) (2015) 937–950.
- [52] B.M. Horton, Noise-modulated distance measuring systems, *Proc. IRE* 47 (5) (1959) 821–828.

- [53] X. Xu, R.M. Narayanan, Range sidelobe suppression technique for coherent ultra wideband random noise radar imaging, *IEEE Trans. Antennas Propag.* 49 (12) (2001) 1836–1842.
- [54] B.C. Flores, E.A. Solis, G. Thomas, Assessment of chaos-based FM signals for range-Doppler imaging, *IEE Proc. Radar Sonar Navig.* 150 (4) (2003) 313–322.
- [55] M. Malanowski, K. Kulpa, Detection of moving targets with continuous-wave noise radar: theory and measurements, *IEEE Trans. Geosci. Remote Sci.* 50 (9) (2012) 3502–3509.
- [56] Q. Yang, Y. Zhang, X. Gu, Design of ultralow sidelobe chaotic radar signal by modulating group delay method, *IEEE Trans. Aerosp. Electron. Syst.* 51 (4) (2015) 3023–3035.
- [57] S.H. Han, J.H. Lee, An overview of peak-to-average power ratio reduction techniques for multicarrier transmission, *IEEE Wirel. Commun.* 12 (2) (2005) 56–65.
- [58] J. Jakabosky, L. Ryan, S.D. Blunt, Transmitter-in-the-loop optimization of distorted OFDM radar emissions, in: IEEE Radar Conf., Ottawa, Canada, Apr./May 2013.
- [59] M.A. Richards, J.A. Scheer, W.A. Holm, *Principles of Modern Radar*, Vol. I: Basic Principles, SciTech, Raleigh, NC, USA, 2010.
- [60] J.J. Kroszczynski, Pulse compression by means of linear-period modulation, *Proc. IEEE* 57 (7) (1969) 1260–1266.
- [61] J. Li, P. Stoica (Eds.), *MIMO Radar Signal Processing*, John Wiley & Sons, Inc., Hoboken, NJ, 2009.
- [62] [www.analog.com/media/cn/training-seminars/tutorials/450968421DDS\\_Tutorial\\_rev12-2-99.pdf](http://www.analog.com/media/cn/training-seminars/tutorials/450968421DDS_Tutorial_rev12-2-99.pdf).
- [63] [www.hit.bme.hu/~papay/edu/Lab/33220A\\_Tutorial.pdf](http://www.hit.bme.hu/~papay/edu/Lab/33220A_Tutorial.pdf).
- [64] [www.tek.com/dl/76W\\_30631\\_0\\_HR\\_Letter.pdf](http://www.tek.com/dl/76W_30631_0_HR_Letter.pdf).
- [65] J.C. Pedro, N.B. Carvalho, *Intermodulation Distortion in Microwave and Wireless Circuits*, Artech House, Boston, MA, 2003.
- [66] B. Cordill, J. Metcalf, S.A. Seguin, D. Chatterjee, S.D. Blunt, The impact of mutual coupling on MIMO radar emissions, in: IEEE Intl. Conf. Electromagnetics in Advanced Applications, Torino, Italy, September 2011.
- [67] G. Babur, P.J. Aubry, F. Le Chevalier, Antenna coupling effects for space-time radar waveforms: analysis and calibration, *IEEE Trans. Antennas Propag.* 62 (5) (2014) 2572–2586.
- [68] J.D. Taylor, *Introduction to Ultra-Wideband Radar Systems*, CRC Press, New York, NY, 1994.
- [69] G.J. Frazer, Y.I. Abramovich, B.A. Johnson, Spatially waveform diverse radar: perspectives for high frequency OTHR, in: IEEE Radar Conf., Boston, MA, April 2007.
- [70] F. Daum, J. Huang, MIMO radar: snake oil or good idea? *IEEE Aerosp. Electron. Syst. Mag.* 24 (5) (2009) 8–12.
- [71] P.M. McCormick, S.D. Blunt, J. Metcalf, Joint spectrum/beampattern design of wideband MIMO radar emissions, in: IEEE Radar Conf., Philadelphia, PA, May 2016.
- [72] P.M. McCormick, S.D. Blunt, J.G. Metcalf, Wideband MIMO frequency-modulated emission design with space-frequency nulling, *IEEE J. Sel. Top. Signal Process.* 11(2) (2017) 363–378.
- [73] <http://www.darpa.mil/program/signal-processing-at-rf>.
- [74] E.R. Billam, Eclipsing effects with high-duty-factor waveforms in long-range radar, *IEE Proc. F Commun. Radar Signal Process.* 132 (7) (1985) 598–603.
- [75] S.D. Blunt, K. Gerlach, E. Mokole, Pulse compression eclipsing repair, in: IEEE Radar Conf., Rome, Italy, May 2008.

- [76] A.M. Klein, M.T. Fujita, Detection performance of hard-limited phase-coded signals, *IEEE Trans. Aerosp. Electron. Syst.* AES-15 (6) (1979) 795–802.
- [77] M.H. Ackroyd, F. Ghani, Optimum mismatched filters for sidelobe suppression, *IEEE Trans. Aerosp. Electron. Syst.* AES-9 (2) (1973) 214–218.
- [78] S.D. Blunt, K. Gerlach, Adaptive pulse compression via MMSE estimation, *IEEE Trans. Aerosp. Electron. Syst.* 42 (2) (2006) 572–584.
- [79] T. Yardibi, J. Li, P. Stoica, M. Xue, A.B. Bagheroer, Source localization and sensing: a nonparametric iterative adaptive approach based on weighted least squares, *IEEE Trans. Aerosp. Electron. Syst.* 46 (1) (2010) 425–443.
- [80] S.D. Blunt, K. Gerlach, T. Higgins, Aspects of radar range super-resolution, in: IEEE Radar Conf., Waltham, MA, April 2007.
- [81] D. Henke, P. McCormick, S.D. Blunt, T. Higgins, Practical aspects of optimal mismatch filtering and adaptive pulse compression for FM waveforms, in: IEEE Intl. Radar Conf., Arlington, VA, May 2015.
- [82] S. Blunt, M. Cook, J. Stiles, Embedding information into radar emissions via waveform implementation, in: Intl. Waveform Diversity & Design Conf., Niagara Falls, Canada, August 2010.
- [83] T. Higgins, S. Blunt, A. Shackelford, Time-range adaptive processing for pulse agile radar, in: Intl. Waveform Diversity & Design Conf., Niagara Falls, Canada, August 2010.
- [84] T. Higgins, K. Gerlach, A. Shackelford, S. Blunt, Aspects of non-identical multiple pulse compression, in: IEEE Radar Conf., Kansas City, MO, May 2011.
- [85] J. Jakabosky, S.D. Blunt, B. Himed, Waveform design and receive processing for non-recurrent nonlinear FMCW radar, in: IEEE Intl. Radar Conf., Arlington, VA, May 2015.
- [86] A. O'Connor, J. Kantor, J. Jakabosky, Joint equalization filters that mitigate waveform-diversity modulation of clutter, in: IEEE Radar Conf., Philadelphia, PA, May 2016.
- [87] A. O'Connor, J. Kantor, J. Jakabosky, Space-time adaptive mismatch processing, in: IEEE Radar Conf., Philadelphia, PA, May 2016.
- [88] J. Jakabosky, S.D. Blunt, B. Himed, Spectral-shape optimized FM noise radar for pulse agility, in: IEEE Radar Conf., Philadelphia, PA, May 2016.
- [89] H.R. Ward, Doppler processor rejection of range ambiguous clutter, *IEEE Trans. Aerosp. Electron. Syst.* AES-11(2) (1975) 519–522.
- [90] J.P. Reilly, Clutter rejection limitations from ambiguous range clutter, in: IEEE Intl. Radar Conf., Arlington, VA, May 1990.
- [91] F.E. Nathanson, J.P. Reilly, M.N. Cohen, *Radar Design Principles*, second ed., SciTech, Raleigh, NC, USA, 1999.
- [92] J.B. Anderson, T. Aulin, C.-E. Sundberg, *Digital Phase Modulation*, Plenum Press, New York, NY, 1986.
- [93] IRIG Standard 106-00: Telemetry Standards, Range Commanders Council Telemetry Group, Range Commanders Council, White Sands Missile Range, New Mexico.
- [94] Bandwidth-Efficient Modulations: Summary of Definitions, Implementation, and Performance, Report Concerning Space Data System Standards, Informational Report CCSDS 413.0-G-2.
- [95] Specifications of the Bluetooth System, Bluetooth Special Interest Group, ver. 1.2, November 2003.
- [96] J. Jakabosky, P. Anglin, M. Cook, S.D. Blunt, J. Stiles, Non-linear FM waveform design using marginal Fisher's information within the CPM framework, in: IEEE Radar Conf., Kansas City, MO, May 2011.

- [97] J. Jakabosky, S.D. Blunt, B. Himed, Optimization of ‘over-coded’ radar waveforms, in: IEEE Radar Conf., Cincinnati, OH, May 2014.
- [98] H.H. Bauschke, J.M. Borwein, On projection algorithms for solving convex feasibility problems, *SIAM Rev.* 38 (3) (1996) 367–426.
- [99] R. Gerchberg, W. Saxton, A practical algorithm for the determination of the phase from image and diffraction plane pictures, *Optik* 35 (1972) 237–246.
- [100] P. Stoica, H. He, J. Li, New algorithms for designing unimodular sequences with good correlation properties, *IEEE Trans. Signal Process.* 57 (4) (2009) 1415–1425.
- [101] L. Jackson, S. Kay, N. Vankayalapati, Iterative method for nonlinear FM synthesis of radar signals, *IEEE Trans. Aerosp. Electron. Syst.* 46 (2) (2010) 910–917.
- [102] S.U. Pillai, K.Y. Li, R. Zheng, B. Himed, Design of unimodular sequences using generalized receivers, in: IEEE Intl. Radar Conf., Arlington, VA, May 2010.
- [103] L.K. Patton, B.D. Rigling, Phase retrieval for radar waveform optimization, *IEEE Trans. Aerosp. Electron. Syst.* 48 (4) (2012) 3287–3302.
- [104] D.R. Fuhrmann, J.P. Browning, M. Rangaswamy, Signaling strategies for the hybrid MIMO phased-array radar, *IEEE J. Sel. Top. Signal Process.* 4 (1) (2010) 66–78.
- [105] J. Jakabosky, P. McCormick, S.D. Blunt, Implementation & design of physical radar waveform diversity, *IEEE AES Syst. Mag.* 31(12) (2016) 26–33.
- [106] C. Baylis, L. Wang, M. Moldovan, J. Martin, H. Miller, L. Cohen, J. De Graaf, Designing transmitters for spectral conformity: power amplifier design issues and strategies, *IET Radar Sonar Navig.* 5 (6) (2011) 681–685.
- [107] J. Hoversten, S. Schafer, M. Roberg, M. Norris, D. Maksimovic, Z. Popovic, Codesign of PA, supply, and signal processing for linear supply-modulated RF transmitters, *IEEE Trans. Microwave Theory Tech.* 60 (6) (2012) 2010–2020.
- [108] M. Fellows, M. Flachsbart, J. Barlow, J. Barkate, C. Baylis, L. Cohen, R.J. Marks, Optimization of power-amplifier load impedance and waveform bandwidth for real-time reconfigurable radar, *IEEE Trans. Aerosp. Electron. Syst.* 51 (3) (2015) 1961–1971.
- [109] J. Barkate, M. Flachsbart, Z. Hays, M. Fellows, J. Barlow, C. Baylis, L. Cohen, R. J. Marks II, Fast, simultaneous optimization of power amplifier input power and load impedance for power-added efficiency and adjacent-channel power ratio using the power Smith tube, *IEEE Trans. Aerosp. Electron. Syst.* 52 (2) (2016) 928–937.
- [110] E. Costa, M. Midrio, S. Pupolin, Impact of amplifier nonlinearities on OFDM transmission system performance, *IEEE Commun. Lett.* 3 (2) (1999) 37–39.
- [111] P. McCormick, J. Jakabosky, S.D. Blunt, C. Allen, B. Himed, Joint polarization/waveform design and adaptive receive processing, in: IEEE Intl. Radar Conf., Arlington, VA, May 2015.
- [112] S.D. Blunt, P. McCormick, T. Higgins, M. Rangaswamy, Physical emission of spatially-modulated radar, *IET Radar Sonar Navig.* 8 (12) (2014) 1234–1246.
- [113] P. McCormick, S.D. Blunt, Fast-time 2-D spatial modulation of physical radar emissions, in: Intl. Radar Symp., Dresden, Germany, June 2015.
- [114] M. Rolfs, Microsaccades: small steps on a long way, *Vision Res.* 49 (2009) 2415–2441.
- [115] E. Ahissar, A. Arieli, Seeing via miniature eye movements: a dynamic hypothesis for vision, *Front. Comput. Neurosci.* 6(89) (2012) 1–27.
- [116] P.M. McCormick, T. Higgins, S.D. Blunt, M. Rangaswamy, Adaptive receive processing of spatially modulated physical radar emissions, *IEEE J. Sel. Top. Signal Process.* 9 (8) (2015) 1415–1426.
- [117] P. Antonik, M.C. Wicks, H.D. Griffiths, C.J. Baker, Frequency diverse array radars, in: IEEE Radar Conf., Verona, NY, April 2006.

- [118] T. Higgins, S. Blunt, Analysis of range-angle coupled beamforming with frequency-diverse chirps, in: Intl. Waveform Diversity & Design Conf., Orlando, FL, February 2009.
- [119] P.F. Sammartino, C.J. Baker, H.D. Griffiths, Frequency diverse MIMO techniques for radar, *IEEE Trans. Aerosp. Electron. Syst.* 49 (1) (2013) 201–222.
- [120] W.-Q. Wang, Range-angle dependent transmit beampattern synthesis for linear frequency diverse arrays, *IEEE Trans. Antennas Propag.* 61 (8) (2013) 4073–4081.
- [121] R. Kinsey, Phased array beam spoiling technique, in: IEEE Intl. Antennas & Propagation Symp., Montreal, QC, Canada, July 1997.
- [122] D.P. Scholnik, A parameterized pattern-error objective for large-scale phase-only array pattern design, *IEEE Trans. Antennas Propag.* 64 (1) (2016) 89–98.
- [123] D. Fuhrmann, G. San Antonio, Transmit beamforming for MIMO radar systems using signal cross-correlation, *IEEE Trans. Aerosp. Electron. Syst.* 44 (1) (2008) 171–186.
- [124] P. Stoica, J. Li, X. Zhu, Waveform synthesis for diversity-based transmit beampattern design, *IEEE Trans. Signal Process.* 56 (6) (2008) 2593–2698.
- [125] H. He, P. Stoica, J. Li, Wideband MIMO systems: signal design for transmit beampattern synthesis, *IEEE Trans. Signal Process.* 59 (2) (2011) 618–628.
- [126] G. Krieger, MIMO-SAR: opportunities and pitfalls, *IEEE Trans. Geosci Remote Sens.* 52 (5) (2014) 2628–2645.
- [127] G.C. Tavik, C.L. Hilterbrick, J.B. Evins, J.J. Alter, J.G. Crnkovich, J.W. de Graaf, W. Habicht, G.P. Hrin, S.A. Lessin, D.C. Wu, S.M. Hagewood, The advanced multi-function RF concept, *IEEE Trans. Microwave Theory Tech.* 53 (3) (2005) 1009–1020.
- [128] H.F. Engler, System considerations for large percent-bandwidth radar, in: IEEE National Telesystems Conf., Atlanta, GA, March 1991.
- [129] M.A. Richards, Fundamentals of Radar Signal Processing, second ed., McGraw-Hill, 2014.

# Geometric foundations for radar signal processing

# 2

**Kevin J. Sangston**

*Georgia Tech Research Institute, Atlanta, GA, United States*

*If you want to improve, be content to be thought foolish and stupid.*

**Epictetus**

*'I will arise and take a chance, too, by my faith!' Nothing ventured, nothing gained, or so men say.*

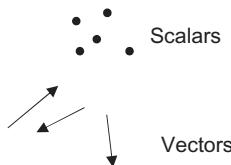
**Geoffrey Chaucer, Canterbury Tales, The Reeves Tale**

---

## 2.1 INTRODUCTION

Geometry forms a foundation for much of what engineers, scientists, and mathematicians do. Hand in hand with algebra, advances in geometry have been at the core of much progress in mathematics and, consequently, science and engineering. However, of all of the geometric notions—i.e., points, lines, planes, spheres, etc.—electrical engineers have most often limited their use primarily to the simplest notions of points and lines. More specifically, electrical engineers work with points—i.e., scalars in the form of real numbers—and lines—i.e., vectors whose components are real numbers. Engineers also commonly work with complex numbers and vectors, which we will say more about later, but generally do not give them geometric interpretations. Even when engineers do interpret them geometrically, it is often in terms of the “direction” of the complex vector, i.e., they interpret the complex vector as a line. Whether such a geometrical interpretation is meaningful is not clear, but complex vectors are primarily used algebraically rather than geometrically, and therefore traditionally the geometric interpretation has been little more than a heuristic aid in discussing problems.

Geometrically, one may think of a scalar as an *unoriented real number*, i.e., a number attached to a point. Because a scalar has no direction, it may be thought of as having no dimension. A common example is temperature. On the other hand, one may think of a vector as a real number with an associated direction. Common examples are forces and velocities, which constitute a magnitude in a specified direction. These notions are illustrated in Fig. 2.1.

**FIG. 2.1**

Geometric view of scalars and vectors.

In particular, a real vector comprises three things:

- (1) A line specifying the direction of the vector;
- (2) A real number specifying the magnitude of the vector; and
- (3) A sign specifying the orientation of the vector along the line.

Thus geometrically a vector may be thought of as a one-dimensional quantity representing an *oriented length*.

It is apparent that the notion of a vector extends the idea of a zero-dimensional number—i.e., a real number treated as a scalar—to a one-dimensional number—i.e., a real number with a direction. It would only seem natural, then that such an extension could be carried to higher dimensions. Just as a scalar is a real number associated with a point, and a vector is a real number associated with a line, it seems one should be able to define a real number associated with a plane and a real number associated with a volume, and so on. Hermann Grassmann pursued just such an extension in 1844—he called the resulting quantity associated with a plane a *bivector*; it represents an *oriented area* (just as a vector represents an oriented length). This notion is illustrated in Fig. 2.2.

He also continued this extension process to higher dimensions and defined an *oriented volume*, an *oriented 4-volume*, etc. In particular, for each  $k=0, 1, 2, \dots$  he defined a  $k$ -vector to represent an oriented number comprising three aspects:

- (1) A  $k$ -dimensional space to specify the “direction” of the  $k$ -vector, i.e., the space that the  $k$ -vector is associated with;
- (2) A real number to specify the magnitude of the  $k$ -vector; and
- (3) A sign to specify the orientation of the  $k$ -vector within the  $k$ -dimensional space.

In Grassmann’s approach, a point is a zero-dimensional space, a line a one-dimensional space, a plane a two-dimensional space, etc., and associated with each, respectively, are 0-vectors (scalars), 1-vectors (vectors), 2-vectors (bivectors), etc.

**FIG. 2.2**

Geometric view of a bivector.

**FIG. 2.3**

Geometric view of a trivector.

For example, the 3-vector associated with a three-dimensional volume is known as a *trivector*, illustrated in Fig. 2.3.

The question now arises as to how electrical engineers can profitably use Grassmann's extended concepts. One way arises from the observation that electrical engineers routinely use complex numbers. For example, due to Euler's formula it is quite convenient to represent sinusoidal signals as complex phasors, which are often described as vectors that rotate in a plane. However, it is not necessarily most effective to think of complex phasors as vectors. For one thing, phasors are necessarily two-dimensional objects (i.e., they lie in a plane) and so geometrically should be associated with a plane rather than a line. Having now learned there is a bivector concept associated with a plane, it should not surprise most readers to learn that complex numbers may be defined in terms of bivectors (as will be discussed later). In addition, complex numbers are often used in engineering problems to implement rotations in a plane. For example, if the complex number  $c_1$  is thought of as a "vector," then multiplying it by another complex number  $c_2$  results in another "vector"  $c_3 = c_2 c_1$ . However, in this formulation the role of the complex number  $c_2$  is fundamentally different from the roles of the "vectors"  $c_1$  and  $c_3$ . In particular, in this expression  $c_2$  is not a "vector" but rather is an *operator* that rotates one "vector" into another "vector." Thus complex numbers can have different attributes depending on the role they play in the problem; it would generally be good to take a systematic approach and keep such differences clear. These are simply two examples, but they suffice to make clear that approaching these notions geometrically can lead to simplifications and clarifications, and electrical engineers may potentially benefit from studying these ideas by bringing geometric intuition to bear on problems using complex numbers.

As another example, consider that in addition to phasor representations engineers also frequently use the concept of a complex vector, i.e., a collection of data samples represented as

$$\begin{bmatrix} \mathcal{J}_1 + i\mathcal{Q}_1 \\ \vdots \\ \mathcal{J}_N + i\mathcal{Q}_N \end{bmatrix}$$

For example,  $\mathcal{J}_1, \dots, \mathcal{J}_N$  might represent the in-phase samples and  $\mathcal{Q}_1, \dots, \mathcal{Q}_N$  the quadrature samples of a signal. Although such constructions lead to algebraic simplifications and thereby can give insight into problem solutions, the notion of a complex vector nonetheless has a certain unsatisfying quality to it. In particular, if one takes the approach discussed before and thinks of a vector as a one-dimensional oriented number—i.e., a number associated with a direction—then the obvious question

is: what is the direction of a *complex* vector? Moreover, if each complex component of the complex vector is itself a vector—i.e., if the phasor concept applies and each complex component is a vector that rotates in a plane—then should a complex vector be interpreted as a vector of vectors? If so, what is its geometric meaning?

To get a taste of the issues involved in addressing this question, consider a real vector defined as follows:

$$\mathbf{r} = \begin{bmatrix} r_1 \\ r_2 \end{bmatrix}$$

In writing a vector in this fashion, often what is understood but not specifically stated is that there are two orthonormal vectors  $\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2$  such that the vector  $\mathbf{r}$  may be written as

$$\mathbf{r} = r_1 \hat{\mathbf{e}}_1 + r_2 \hat{\mathbf{e}}_2$$

This vector is illustrated in Fig. 2.4.

As is well known, one may talk about both the length  $|\mathbf{r}|$  of the vector  $\mathbf{r}$  and its direction  $\theta$ :

$$|\mathbf{r}| = \sqrt{r_1^2 + r_2^2}$$

$$\cos \theta = \frac{r_1}{\sqrt{r_1^2 + r_2^2}}$$

$$\sin \theta = \frac{r_2}{\sqrt{r_1^2 + r_2^2}}$$

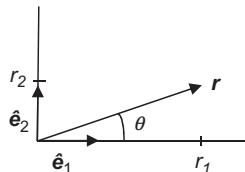
Here the direction is specified in terms of an angle  $\theta$  with respect to the basis vector  $\hat{\mathbf{e}}_1$ . Thus geometrically the direction of the vector  $\mathbf{r}$  is well defined.

However, consider the same situation except now instead of being real the vector  $\mathbf{c}$  is complex:

$$\mathbf{c} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} c_{1r} + ic_{1i} \\ c_{2r} + ic_{2i} \end{bmatrix}$$

It is common to define the “length” of this complex vector as

$$|\mathbf{c}| = \sqrt{\mathbf{c}^H \mathbf{c}}$$



**FIG. 2.4**

Components of a vector.

where

$$\mathbf{c}^H = [c_1^* \ c_2^*] = [c_{1r} - ic_{1i} \ c_{2r} - ic_{2i}]$$

Carrying out this computation yields

$$|\mathbf{c}| = \sqrt{c_1^* c_1 + c_2^* c_2} = \sqrt{c_{1r}^2 + c_{1i}^2 + c_{2r}^2 + c_{2i}^2}$$

What about the “direction” of a complex vector? Heuristically one often finds references to directional concepts in relation to complex vectors. For example, in radar processing it is quite common to encounter complex “steering” vectors, a concept that inherently assumes that the “direction” of a complex vector is a meaningful concept. Thus one should understand how to define the “direction” of  $\mathbf{c}$ . There are likely several ways to approach this problem, but one clue to a possible approach lies in the fact that the magnitude of  $\mathbf{c}$  is determined by the four quantities  $c_{1r}, c_{1i}, c_{2r}, c_{2i}$ . This observation suggests representing the two-dimensional complex vector  $\mathbf{c}$  by a *four-dimensional real* vector  $\mathbf{d}$ :

$$\mathbf{d} = \begin{bmatrix} c_{1r} \\ c_{2r} \\ c_{1i} \\ c_{2i} \end{bmatrix}$$

The length of  $\mathbf{d}$  in this four-dimensional space is naturally defined to be

$$|\mathbf{d}| = \sqrt{c_{1r}^2 + c_{1i}^2 + c_{2r}^2 + c_{2i}^2} = |\mathbf{c}|$$

Thus the “length” of  $\mathbf{c}$  and  $\mathbf{d}$  is the same. Now let the unit vectors  $\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \hat{\mathbf{e}}_3, \hat{\mathbf{e}}_4$  form an orthonormal basis for this four-dimensional space and write this vector as

$$\mathbf{d} = c_{1r}\hat{\mathbf{e}}_1 + c_{2r}\hat{\mathbf{e}}_2 + c_{1i}\hat{\mathbf{e}}_3 + c_{2i}\hat{\mathbf{e}}_4$$

With this formulation, it is possible to specify the direction of the four-dimensional real vector  $\mathbf{d}$  relative to a chosen unit vector, say  $\hat{\mathbf{e}}_1$ , by specifying appropriate angles in the  $\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2$  plane, the  $\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_3$  plane, and the  $\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_4$  plane. Hence, by treating the two-dimensional complex vector  $\mathbf{c}$  as a four-dimensional real vector  $\mathbf{d}$ , it is possible to define both the length and the direction of  $\mathbf{c}$ .

However, in what sense are the two-dimensional complex vector  $\mathbf{c}$  and the four-dimensional real vector  $\mathbf{d}$  actually the “same” vector? To explore this question, observe that because  $\mathbf{d}$  is a real vector it is possible to write

$$\mathbf{d} = \begin{bmatrix} c_{1r} \\ c_{2r} \\ c_{1i} \\ c_{2i} \end{bmatrix} = c_{1r}\hat{\mathbf{e}}_1 + c_{1i}\hat{\mathbf{e}}_3 + c_{2r}\hat{\mathbf{e}}_2 + c_{2i}\hat{\mathbf{e}}_4$$

The order of terms on the right-hand side has been rearranged for convenience. It will prove interesting now to define the following products:

$$\hat{\mathbf{e}}_1\hat{\mathbf{e}}_1 = 1$$

$$\hat{\mathbf{e}}_2\hat{\mathbf{e}}_2 = 1$$

These definitions will be examined more closely later, but for now consider them simply as formal definitions. With these definitions, it is now possible to factor out the unit vectors  $\hat{\mathbf{e}}_1$  and  $\hat{\mathbf{e}}_2$ :

$$c_{1r}\hat{\mathbf{e}}_1 + c_{1i}\hat{\mathbf{e}}_3 = (c_{1r} + c_{1i}\hat{\mathbf{e}}_3\hat{\mathbf{e}}_1)\hat{\mathbf{e}}_1$$

$$c_{2r}\hat{\mathbf{e}}_2 + c_{2i}\hat{\mathbf{e}}_4 = (c_{2r} + c_{2i}\hat{\mathbf{e}}_4\hat{\mathbf{e}}_2)\hat{\mathbf{e}}_2$$

Like  $\hat{\mathbf{e}}_1\hat{\mathbf{e}}_1$  and  $\hat{\mathbf{e}}_2\hat{\mathbf{e}}_2$ , the objects  $\hat{\mathbf{e}}_3\hat{\mathbf{e}}_1$  and  $\hat{\mathbf{e}}_4\hat{\mathbf{e}}_2$  will be studied further later, but for now they may be viewed as simply objects that are being formally manipulated. The vector  $\mathbf{d}$  becomes

$$\mathbf{d} = (c_{1r} + c_{1i}\hat{\mathbf{e}}_3\hat{\mathbf{e}}_1)\hat{\mathbf{e}}_1 + (c_{2r} + c_{2i}\hat{\mathbf{e}}_4\hat{\mathbf{e}}_2)\hat{\mathbf{e}}_2$$

By analogy with the real vector  $\mathbf{r}$  considered before, this relation now suggests writing

$$\mathbf{d} = (c_{1r} + c_{1i}\hat{\mathbf{e}}_3\hat{\mathbf{e}}_1)\hat{\mathbf{e}}_1 + (c_{2r} + c_{2i}\hat{\mathbf{e}}_4\hat{\mathbf{e}}_2)\hat{\mathbf{e}}_2 = \begin{bmatrix} c_{1r} + \hat{\mathbf{e}}_3\hat{\mathbf{e}}_1 c_{1i} \\ c_{2r} + \hat{\mathbf{e}}_4\hat{\mathbf{e}}_2 c_{2i} \end{bmatrix}$$

This same analogy also suggests writing

$$\mathbf{c} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = c_1\hat{\mathbf{e}}_1 + c_2\hat{\mathbf{e}}_2$$

A comparison of these formal representations of  $\mathbf{c}$  and  $\mathbf{d}$  suggests that if  $\hat{\mathbf{e}}_3\hat{\mathbf{e}}_1$  and  $\hat{\mathbf{e}}_4\hat{\mathbf{e}}_2$  can be understood to act *algebraically* as the unit imaginary number  $i$  then it is possible to write

$$c_1 = c_{1r} + c_{1i}\hat{\mathbf{e}}_3\hat{\mathbf{e}}_1$$

$$c_2 = c_{2r} + c_{2i}\hat{\mathbf{e}}_4\hat{\mathbf{e}}_2$$

In this way it is possible to equate the four-dimensional *real* vector  $\mathbf{d}$  to the two-dimensional *complex* vector  $\mathbf{c}$ .

$$\mathbf{d} = \mathbf{c}$$

Thus in this sense  $\mathbf{d}$  and  $\mathbf{c}$  are the same vector.

As has been shown, it is possible to give a geometric interpretation in terms of length and direction to a complex vector. As will be discussed further later, this interpretation yields an alternative, geometric understanding of various analyses with complex vectors in the context of signal processing problems. For example, a complex vector formed from the in-phase and quadrature samples of a sinusoidal signal at a given frequency may be viewed as a real vector confined to a two-dimensional subspace, i.e., a plane. Moreover, this plane is determined by the frequency of the signal. Thus if this complex vector is multiplied by a complex constant, the result is to change the length of the vector and to rotate it within the plane by an amount determined by the phase of the complex constant. It is a common problem in detection processing that one often knows a complex steering vector of interest, but the data vector includes an unknown complex constant. To match this data properly to the steering vector, one would need to know the phase of this complex constant and rotate the steering vector

within the appropriate plane by an amount determined by this phase. One could then project the data vector onto the rotated steering vector to assess how well the two vectors match. However, if the phase of the steering vector is unknown, which is a common situation, then projecting the data vector onto the steering vector is not very helpful since the data vector and the steering vector are misaligned due to the unknown phase. In a sense, a complex steering vector in this case is better viewed as a real steering plane. Then one can project the data vector onto the plane within which the associated real steering vector rotates. In doing so, one in effect implements a matched subspace detector rather than a matched filter detector. This observation explains why it is meaningful to compare the magnitude of a matched filter output to a threshold to make a detection decision when the phase is unknown. The magnitude of the matched filter output when the phase is unknown turns out to be the length of the data vector that is orthogonally projected onto the plane of the steering vector. It is only when the phase of the multiplicative constant is known that one can meaningfully project this vector further onto the steering vector itself. Without the ability to project onto the appropriately rotated steering vector, the best one can do to assess how close the data vector is to the steering plane—and hence how close it is to having the frequency associated with the plane—is to examine the length of this orthogonally projected vector. This and other examples will be discussed further in the pages that follow.

One final observation by way of motivation is in order. Signal processing engineers often think in terms of subspaces and projections of vectors into subspaces. Consequently, it would perhaps prove fruitful to have an *algebra of subspaces*. To some extent the traditional matrix-based approach to linear algebra provides an algebraic approach to working with subspaces. However, Grassmann's exterior algebra provides a more natural setting for developing an algebra of subspaces and finds a very convenient formulation in terms of geometric algebra.

---

## 2.2 GEOMETRIC ALGEBRA

### 2.2.1 HOW TO MULTIPLY VECTORS

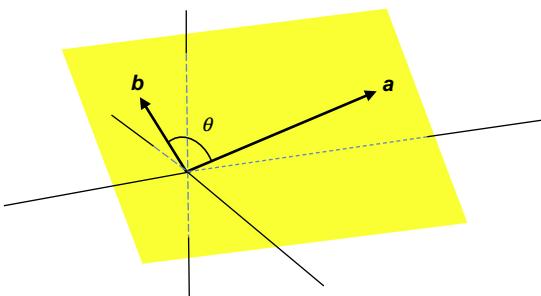
The basic algebraic operations useful to engineers are addition and multiplication. Of course, this idea also encompasses both subtraction and division, which are simply addition and multiplication with inverse elements. In working with vectors, one encounters no problem with addition, which extends from real scalars to real vectors in a straightforward way. However, the situation is somewhat different with multiplication. Generally speaking, most engineers (unless they have an unusually deep background in mathematics) do not know how to multiply vectors. The goal of this section is to explore how to do so.

#### 2.2.1.1 A Nonassociative Product of Vectors

Consider two vectors in a three-dimensional Euclidean<sup>1</sup> space as shown in Fig. 2.5.

---

<sup>1</sup>Although much (if not all) of what follows is valid (or can be made valid) for more general non-Euclidean spaces, such generality is not needed and so attention is restricted to Euclidean spaces.

**FIG. 2.5**

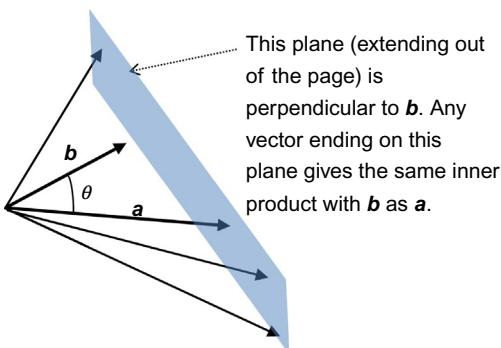
Geometric relation between two vectors.

The goal is to define multiplication *and* division of these two vectors. Most engineers are familiar with at least two ways to “multiply” vectors. First there is the inner product

$$\mathbf{a} \cdot \mathbf{b} = \mathbf{b} \cdot \mathbf{a} = |\mathbf{a}| |\mathbf{b}| \cos \theta. \quad (2.1)$$

In writing the inner product in this way coordinates have not been used. This is because the choice of the axes of a coordinate system to represent the vectors is irrelevant to the inner product. Although coordinates can be useful in computing the inner product, *geometrically* the inner product depends only on the magnitudes of the two vectors and the smaller of the two angles between them in their shared plane. These quantities are the same in any coordinate system.

Observe now that the inner product is *not invertible*. Given the inner product  $\mathbf{a} \cdot \mathbf{b}$  and one of the vectors, e.g.,  $\mathbf{b}$ , one cannot obtain the other vector  $\mathbf{a}$ , as may be seen from Fig. 2.6 (in which  $\mathbf{b}$  lies in the plane of the page). Thus the inner product cannot be used by itself to define multiplication *and* division of vectors. The inner product does not represent multiplication in the full algebraic sense of interest here.

**FIG. 2.6**

Inner product is not invertible.

In addition to the inner product, in a *three-dimensional space* engineers are also generally familiar with the cross product:

$$\mathbf{a} \times \mathbf{b} = |\mathbf{a}| |\mathbf{b}| \sin \theta \hat{\mathbf{n}} \quad (2.2)$$

where  $\hat{\mathbf{n}}$  is a unit vector pointing normal to the plane defined by  $\mathbf{a}$  and  $\mathbf{b}$  and determined in accordance with a right-handed coordinate system. Again, this formulation does not depend on a specific coordinate system for the three-dimensional space other than to require right-handedness to determine the orientation of  $\hat{\mathbf{n}}$ . Like the inner product, the cross product is also not invertible. Thus it cannot be used it by itself to define both multiplication and division.

The first question that arises from these considerations is whether a true algebraically invertible product of vectors can be defined at all. To explore this question, observe first that if *both* the inner and cross products as well as one of the vectors are available, then this information suffices to determine the other vector. To see this consider the basic relations

$$\mathbf{a} \cdot \mathbf{b} = |\mathbf{a}| |\mathbf{b}| \cos \theta \quad (2.3)$$

and

$$|\mathbf{a} \times \mathbf{b}| = |\mathbf{a}| |\mathbf{b}| \sin \theta. \quad (2.4)$$

From these relations it follows that:

$$|\mathbf{a}| |\mathbf{b}| = \left[ (\mathbf{a} \cdot \mathbf{b})^2 + (|\mathbf{a} \times \mathbf{b}|)^2 \right]^{1/2} \quad (2.5)$$

and

$$\theta = \tan^{-1} \frac{|\mathbf{a} \times \mathbf{b}|}{\mathbf{a} \cdot \mathbf{b}}. \quad (2.6)$$

Therefore from these two products and one of the vectors, say  $\mathbf{b}$ , one can compute the magnitude  $|\mathbf{a}|$  of the other vector. Then the angle between the two vectors, which satisfies  $0 \leq \theta \leq \pi$ , the plane they are in (determined by  $\hat{\mathbf{n}}$ ), and their relative orientation (from the assumption of a right-handed coordinate system) is enough information to determine the complete vector  $\mathbf{a}$ .

This brief examination of the geometric relations inherent in the inner and cross products shows that vectors can be used for division in the sense that one can obtain one of the vectors from these two products and the other vector. The question now arises as to whether this division can be defined *algebraically*, i.e., can  $\mathbf{b}$  be obtained from knowledge of  $\mathbf{a} \cdot \mathbf{b}$ ,  $\mathbf{a} \times \mathbf{b}$ , and  $\mathbf{a}$  with simple algebraic manipulations. To begin exploring this question define a product of two vectors  $\mathbf{a}$  and  $\mathbf{b}$  by the relation

$$\mathbf{a} \odot \mathbf{b} = \mathbf{a} \cdot \mathbf{b} + \mathbf{a} \times \mathbf{b}. \quad (2.7)$$

As part of this product also define

$$\lambda \odot \mathbf{b} = \mathbf{b} \odot \lambda = \lambda \mathbf{b} \quad (2.8)$$

where  $\lambda$  is an arbitrary  $\odot$ scalar (i.e., real number) and  $\mathbf{b}$  is an arbitrary vector. In this approach the product of two vectors is neither a scalar alone nor a vector alone but

rather is an object comprising the sum of a scalar and a vector. Engineers do not generally deal with such objects, but just as the sum of a real number and an imaginary number, which are different kinds of objects, is routinely used, one can likewise expand the class of objects of interest to allow for the sum of real numbers and vectors.

From the property that the cross product of a vector with itself is equal to zero, an arbitrary vector  $\mathbf{a}$  satisfies

$$\mathbf{a}^2 = \mathbf{a} \odot \mathbf{a} = \mathbf{a} \cdot \mathbf{a} = |\mathbf{a}|^2 \quad (2.9)$$

Note that this quantity is a scalar. Thus in this approach the square of a vector is a scalar. This idea will prove very important in the material that follows. As a result of  $\mathbf{a}^2$  being a scalar, one can define the *inverse vector* of any nonzero vector  $\mathbf{a}$  as

$$\mathbf{a}^{-1} = \frac{1}{\mathbf{a}} \triangleq \frac{\mathbf{a}}{\mathbf{a}^2} = \frac{\mathbf{a}}{|\mathbf{a}|^2} \quad (2.10)$$

By inverse vector of  $\mathbf{a}$  is meant a vector  $\mathbf{a}^{-1}$  that satisfies

$$\mathbf{a}^{-1} \odot \mathbf{a} = \mathbf{a} \odot \mathbf{a}^{-1} = 1 \quad (2.11)$$

Note that with the definition in Eq. (2.9),  $\mathbf{a}^{-1}$  is a scalar times a vector and hence is a vector. It immediately follows from the definition in Eq. (2.9) that

$$\mathbf{a}^{-1} \odot \mathbf{a} = \frac{\mathbf{a}}{|\mathbf{a}|^2} \odot \mathbf{a} = \frac{|\mathbf{a}|^2}{|\mathbf{a}|^2} = 1 \quad (2.12)$$

Similarly, one finds that  $\mathbf{a} \odot \mathbf{a}^{-1} = 1$ . Thus  $\mathbf{a}^{-1}$  is the desired inverse vector. Now examine

$$(\mathbf{a} \odot \mathbf{b}) \odot \mathbf{a}^{-1} = \frac{(\mathbf{a} \odot \mathbf{b}) \odot \mathbf{a}}{|\mathbf{a}|^2} = \frac{1}{|\mathbf{a}|^2} (\mathbf{a} \cdot \mathbf{b} + \mathbf{a} \times \mathbf{b}) \odot \mathbf{a} \quad (2.13)$$

Require that the  $\odot$  product of vectors be distributive over addition and write

$$(\mathbf{a} \odot \mathbf{b}) \odot \mathbf{a}^{-1} = \frac{1}{|\mathbf{a}|^2} [(\mathbf{a} \cdot \mathbf{b}) \odot \mathbf{a} + (\mathbf{a} \times \mathbf{b}) \odot \mathbf{a}] \quad (2.14)$$

The first term on the right-hand side is the  $\odot$  product of the scalar  $\frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{a}|^2}$  with the vector  $\mathbf{a}$ . This is given by the usual product of a scalar and a vector. Thus this first term is a vector:

$$\frac{(\mathbf{a} \cdot \mathbf{b})}{|\mathbf{a}|^2} \odot \mathbf{a} = \frac{(\mathbf{a} \cdot \mathbf{b})}{|\mathbf{a}|^2} \mathbf{a} \quad (2.15)$$

Apply the  $\odot$  product again to the second term on the right-hand side to obtain

$$(\mathbf{a} \odot \mathbf{b}) \odot \mathbf{a}^{-1} = \frac{1}{|\mathbf{a}|^2} [(\mathbf{a} \cdot \mathbf{b}) \mathbf{a} + (\mathbf{a} \times \mathbf{b}) \cdot \mathbf{a} + (\mathbf{a} \times \mathbf{b}) \times \mathbf{a}] \quad (2.16)$$

Observe now that  $\mathbf{a} \times \mathbf{b}$  is perpendicular to both  $\mathbf{a}$  and  $\mathbf{b}$ . Therefore  $(\mathbf{a} \times \mathbf{b}) \cdot \mathbf{a} = 0$  and

$$(\mathbf{a} \odot \mathbf{b}) \odot \mathbf{a}^{-1} = \frac{1}{|\mathbf{a}|^2} [(\mathbf{a} \cdot \mathbf{b})\mathbf{a} + (\mathbf{a} \times \mathbf{b}) \times \mathbf{a}] \quad (2.17)$$

Use the triple product rule of the cross product on the last term to write

$$(\mathbf{a} \times \mathbf{b}) \times \mathbf{a} = -\mathbf{a} \times (\mathbf{a} \times \mathbf{b}) = -\mathbf{a}(\mathbf{a} \cdot \mathbf{b}) + \mathbf{b}(\mathbf{a} \cdot \mathbf{a}) \quad (2.18)$$

This result yields

$$(\mathbf{a} \odot \mathbf{b}) \odot \mathbf{a}^{-1} = \frac{1}{|\mathbf{a}|^2} [(\mathbf{a} \cdot \mathbf{b})\mathbf{a} - \mathbf{a}(\mathbf{a} \cdot \mathbf{b}) + \mathbf{b}(\mathbf{a} \cdot \mathbf{a})] = \mathbf{b} \quad (2.19)$$

The  $\odot$  product thus allows one to perform vector multiplication *and* division *algebraically*!

Even though most engineers do not think in terms of dividing by vectors, such an operation is certainly possible as the previous considerations show. Showing that it is possible, however, is not sufficient, as the algebraic product of two vectors must also be useful and widely applicable. The  $\odot$  product defined before has at least two significant failings that make it difficult to use in a general setting. First, the cross product is not defined for all dimensions and thus the  $\odot$  product would not be available in all dimensions. Most signal processing problems are not limited, for example, to three dimensions, and so the  $\odot$  product defined herein would not be very useful for signal processing applications. Second, the cross product is not associative, i.e., in general  $(\mathbf{a} \times \mathbf{b}) \times \mathbf{c} \neq \mathbf{a} \times (\mathbf{b} \times \mathbf{c})$ . Therefore the  $\odot$  product is also not associative, i.e.

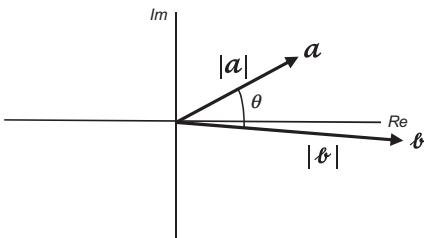
$$(\mathbf{a} \odot \mathbf{b}) \odot \mathbf{c} \neq \mathbf{a} \odot (\mathbf{b} \odot \mathbf{c}). \quad (2.20)$$

As a result, in using this product one would have to be very careful to keep track of parentheses, which could become quite tedious. Thus even though the approach sketched before leads to an invertible product, it cannot be used in all dimensions and may be difficult to apply even in three dimensions where it could be used.

Before considering other approaches to defining a suitable product of vectors, one should pause to consider an important point revealed by the previous sketch. Namely, with  $\odot$  product the result of multiplying two vectors is *not a vector*. Instead it is the sum of a scalar and a vector. Therefore to have a meaningful algebraic approach to multiplying and dividing vectors, it seems one must allow for objects that are not limited to scalars and vectors alone, but might include such objects as the sum of scalars and vectors. Moreover, to be able to multiply more than two vectors, one might expect even more objects in the resulting algebra.

### 2.2.1.2 An Associative Product of Vectors

To explore developing an associative product of vectors in any number of dimensions, begin by first restricting attention to two dimensions and then to three dimensions to see if these special cases teach anything that would be helpful for developing an approach to multiplying and dividing vectors. Fig. 2.7 presents the two vectors above, but now interprets them as complex numbers  $a$  and  $b$  in a plane. In this

**FIG. 2.7**

Vector associated with a complex number.

representation of the 2D vectors the complex number associated with vector  $\mathbf{a}$  is given by

$$\mathbf{a} = |\mathbf{a}| e^{i\phi} \quad (2.21)$$

where  $|\mathbf{a}|$  is the magnitude of the vector and  $\phi$  is the angle from the real axis to the vector.

With this representation, the product of two complex numbers is given by

$$\mathbf{a}^* \mathbf{b} = |\mathbf{a}| |\mathbf{b}| \cos \theta + i |\mathbf{a}| |\mathbf{b}| \sin \theta \quad (2.22)$$

where

$$\theta = \phi_b - \phi_a \quad (2.23)$$

In this expression,  $\theta$  is the angle between the two vectors. Note that this result incorporates the definition of an inner product.

To use this formulation, now define the product of two vectors by

$$\mathbf{ab} = \mathbf{a}^* \mathbf{b} = |\mathbf{a}| |\mathbf{b}| \cos \theta + i |\mathbf{a}| |\mathbf{b}| \sin \theta \quad (2.24)$$

The first observation to make is that this product is invertible and thus one can divide by vectors. To see this simply observe that multiplying  $\mathbf{a}^* \mathbf{b}$  from the left by  $\mathbf{a}/|\mathbf{a}|^2$  recovers  $\mathbf{b}$  and hence  $\mathbf{b}$ . In this approach, however, multiplication is *not commutative*:

$$\mathbf{ba} = \mathbf{b}^* \mathbf{a} = |\mathbf{a}| |\mathbf{b}| \cos \theta - i |\mathbf{a}| |\mathbf{b}| \sin \theta \neq \mathbf{ab}. \quad (2.25)$$

As a result it is necessary to define right- and left-multiplication (and hence division) differently. However, this complication should not cause any insurmountable problems because engineers are already accustomed to matrix multiplication, which is also noncommutative.

One defect of this approach is that in the process of computing the product  $\mathbf{ab}$  it was necessary to assign real and imaginary axes. But these axes are irrelevant! The product as defined only depends on the magnitudes of the vectors and the angle *between* them, which does not require a real axis to be assigned as a reference from which to measure angles. For example, rotating these vectors but leaving their magnitudes and the angle between them unchanged leads to the same result for this product even though the corresponding complex numbers change. The product therefore measures the *relative directions* of the vectors (as well as the relative magnitudes), and assigning axes should be unnecessary.

One result of assigning real and imaginary axes has been to introduce  $i = \sqrt{-1}$  into the result. Observe now that  $i$  serves multiple purposes. In particular,

- (1) It defines the imaginary axis and in so doing keeps the “real” part and the “imaginary” part of the complex number separate;
- (2) It defines an orientation in the plane:  $+i\phi$  indicates  $\phi$  is measured from the real axis in a counterclockwise rotational direction, whereas  $-i\phi$  indicates  $\phi$  is measured from the real axis in a clockwise rotational direction;
- (3) It causes a rotation through multiplication; and
- (4) It turns one kind of quantity into another through multiplication, i.e., multiplication of a real number by  $i$  yields an imaginary number, whereas multiplication of an imaginary number by  $i$  yields a real number.

As shall be shown later, the generalization to obtain a vector product suitable for any number of dimensions incorporates these different functions of  $i$  in different ways.

Consider the second function of  $i$ : defining an orientation. It is well known that in writing a complex number as

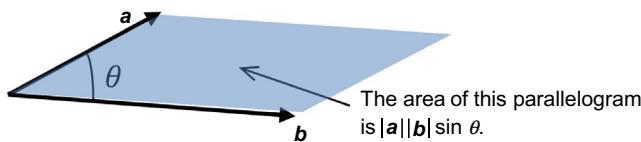
$$\mathbf{a} = |\mathbf{a}|e^{i\phi} \quad (2.26)$$

the role of  $i$  is to cause the angle  $\phi$  to be measured from the real axis in a counterclockwise direction, whereas  $-i$  causes the angle  $\phi$  to be measured in a clockwise direction. From this point of view,  $i$  assigns an orientation to the quantity it multiplies. Now observe that the first term in our product is simply the inner product

$$\mathbf{a} \cdot \mathbf{b} = |\mathbf{a}||\mathbf{b}|\cos\theta. \quad (2.27)$$

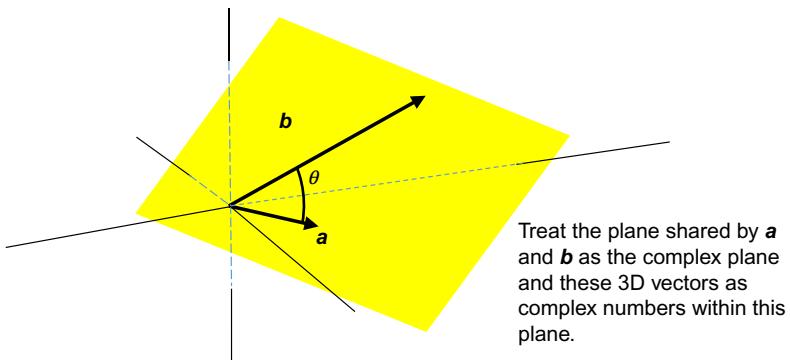
The second term has two parts, namely,  $i$  and  $|\mathbf{a}||\mathbf{b}|\sin\theta$ . The role of  $i$  has already been discussed so now examine  $|\mathbf{a}||\mathbf{b}|\sin\theta$ . As may be seen from Fig. 2.8, it defines an area in the plane containing  $\mathbf{a}$  and  $\mathbf{b}$  (which in this case is the only possible plane since there are only two dimensions). Note that  $|\mathbf{a}||\mathbf{b}|\sin\theta$  is just a *number*—it does not define any particular *shape*, but rather only defines an area. It is convenient to represent this area as the area defined by the parallelogram obtained by sweeping one vector across the other, but this area could just as easily be represented as a circle or some other shape with the same area. What is important for defining the product is that it gives a number representing an area.

The notion that  $i$  assigns an orientation to the quantity it multiplies suggests that the quantity  $i|\mathbf{a}||\mathbf{b}|\sin\theta$  might be fruitfully viewed as an *oriented area*. The resulting product of two vectors would then be the sum of two different kinds of



**FIG. 2.8**

Area of a parallelogram.

**FIG. 2.9**

Geometric relation between two vectors in three dimensions.

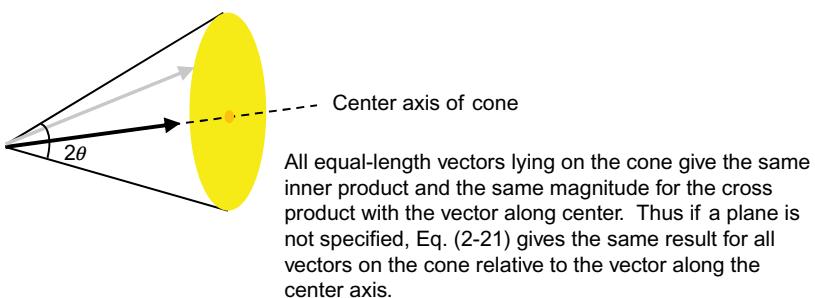
objects: a scalar—given by the inner product—and an oriented area. This notion will be explored more fully later.

First, however, consider extending this approach to three dimensions as shown in Fig. 2.9.

In particular, by analogy with the two-dimensional case formulate a “complex number” to represent the product  $\mathbf{ab}$  as

$$|\mathbf{a}||\mathbf{b}|\cos\theta + i_p|\mathbf{a}||\mathbf{b}|\sin\theta \quad (2.28)$$

where  $i_p$  is to be interpreted as restricting this “complex number” to the plane defined by  $\mathbf{a}$  and  $\mathbf{b}$ . In this way one can use the same approach to defining the product of vectors in three dimensions as was discussed before in two dimensions. Note, however,  $i_p$  cannot simply be  $i = \sqrt{-1}$  since at a minimum it must also *define the plane* if the product is to be invertible. If a plane is not specified, then as shown in Fig. 2.10 two vectors can define a cone with a cone angle equal to  $2\theta$  where one of the vectors is directed along the center axis of the cone and the other vector lies on the cone. In

**FIG. 2.10**

Need to define a plane.

that case the product would not be uniquely invertible. The quantity  $i_p$  should also act like  $i$  by defining a direction in the plane so that vectors can be rotated.

To this end, combine  $i$ —which will give rotations in the plane through multiplication—with the full cross product—which both defines the plane and an orientation in the plane in accordance with the right-hand rule. Therefore setting

$$i_p = i\hat{n} \quad (2.29)$$

yields for the product of two vectors:

$$\mathbf{ab} = \mathbf{a} \cdot \mathbf{b} + i\mathbf{a} \times \mathbf{b}. \quad (2.30)$$

This approach also leads to a product of vectors that can be used to define division. In particular,

$$\mathbf{a}^2 = \mathbf{aa} = \mathbf{a} \cdot \mathbf{a} = |\mathbf{a}|^2 \quad (2.31)$$

i.e., the square of any vector is a scalar equal to the square of its length (this follows since  $\mathbf{a} \times \mathbf{a} = 0$ ). With this notion, again define the inverse of a nonzero vector as

$$\mathbf{a}^{-1} = \frac{1}{\mathbf{a}} = \frac{\mathbf{a}}{\mathbf{a}^2} = \frac{\mathbf{a}}{|\mathbf{a}|^2}. \quad (2.32)$$

It turns out that the product in Eq. (2.30) is associative— $(\mathbf{ab})\mathbf{c} = \mathbf{a}(\mathbf{bc})$ —as may be easily shown (the reader should do this computation). Thus one can invert this product (i.e., divide) to recover  $\mathbf{b}$  by left-multiplying by  $\mathbf{a}/|\mathbf{a}|^2$ :

$$\frac{\mathbf{ab}}{\mathbf{a}} = \mathbf{a}^{-1}\mathbf{ab} = \frac{\mathbf{a}}{|\mathbf{a}|^2}\mathbf{ab} = \frac{\mathbf{a}^2}{|\mathbf{a}|^2}\mathbf{b} = \mathbf{b}. \quad (2.33)$$

Similarly one can recover  $\mathbf{a}$  from the product by right-multiplying by  $\mathbf{b}/|\mathbf{b}|^2$ .

The upshot of the previous discussions is that it is possible to form an invertible, associative product of vectors in two and three dimensions, and therefore one might expect to form such a product in arbitrary dimensions. Note, however, that the goal in going from two dimensions to three dimensions was to *extend* the product that was defined in two dimensions. Unfortunately, the product defined for three dimensions is *not* an extension of the product defined for two dimensions—it does not make sense to formulate in three dimensions an approach that cannot even be defined in two dimensions! In particular, the cross product is not defined in two dimensions as there is simply not a third dimension for the vector defined by the cross product to point in. Moreover, in dimensions higher than three, there is not a unique direction for the vector to point in since there will be multiple directions that are perpendicular to the plane defined by  $\mathbf{a}$  and  $\mathbf{b}$ . Thus despite the formal similarity between the two-dimensional approach based on complex numbers and the three-dimensional approach based on the inner product and the “complex” cross product, it does not appear to be fruitful to use the  $i\mathbf{a} \times \mathbf{b}$  for defining a product of vectors that will work in any number of dimensions.

Using complex numbers to define a product of vectors in two dimensions led to another complex number for the product, and thus apparently the product of two vectors yields another vector in that approach. However, in three dimensions the product

of two vectors led to the sum of the inner product, which is a scalar, and a complex version of the cross product, which is evidently a complex vector that differs in kind from the two vectors that went into the product. Thus the previously defined product of vectors in three dimensions leads to a collection of three different kinds of objects: scalars, vectors, and the complex vectors that arise from the “complexified” cross product. It should not be surprising therefore that in generalizing the product of vectors to any number of dimensions one might find:

- (1) The view of what happens in two dimensions must be changed to allow for the product of vectors to be a sum of different kinds of objects; and
- (2) This view of the product of vectors as a sum of different kinds of objects carries over to any number of dimensions.

In both two and three dimensions the product of the two vectors is the inner product added to something else, where the nature of this “something else” depends on the dimension. In two dimensions this additional quantity is  $i|\mathbf{a}||\mathbf{b}|\sin\theta$  and in three dimensions it is the quite different object  $i\mathbf{a} \times \mathbf{b}$ . It is interesting to explore both how these objects differ and how they are alike.

First consider the differences. Recall from before that  $i$  performs multiple functions in two dimensions and that these functions might be incorporated into higher dimensions in different ways. The first difference manifests in the attempt to go from two to three dimensions. In three dimensions,  $i$  is not needed to distinguish between different kinds of objects—the vector  $\mathbf{a} \times \mathbf{b}$  differs in kind from the scalar  $\mathbf{a} \cdot \mathbf{b}$  without needing  $i$  to distinguish them. Thus the first function of  $i$  in two dimensions, namely, keeping the real and imaginary parts of the complex number separate, is unnecessary in three dimensions.

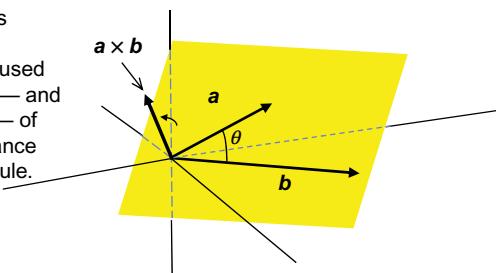
In addition, in three dimensions  $i$  does not need to define a sense of orientation in the plane, since that role is played by the vector  $\hat{\mathbf{n}}$  normal to the plane in conjunction with the right-hand rule as illustrated in Fig. 2.11. Unlike in two dimensions, a plane in three dimensions has two sides and so one must define which side of the plane is being used in conjunction with a rule such as the right-hand rule to orient rotations.

In at least one sense, however, the object  $i\mathbf{a} \times \mathbf{b}$  taken as a whole in three dimensions is similar to  $i|\mathbf{a}||\mathbf{b}|\sin\theta$  in two dimensions in that they both define an *oriented area in the plane defined by  $\mathbf{a}$  and  $\mathbf{b}$* . Just as there are two orientations in the two-dimensional plane according to  $i$  or  $-i$ , there are also two orientations in three dimensions according to  $\mathbf{a} \times \mathbf{b}$  or  $-\mathbf{a} \times \mathbf{b}$  ( $= \mathbf{b} \times \mathbf{a}$ ). It is this observation that suggests a way forward for extending this approach for multiplying vectors to any number of dimensions.

### 2.2.1.3 The Geometric Product of Vectors

Encouraged by the possibility of dividing vectors, one may now ask whether it is possible to use the lessons learned before to develop an *invertible and associative product of vectors* that works in *any number of dimensions*. In particular, the previous considerations demonstrated that if the product of vectors

This normal vector is defined by the cross product and can be used to determine a side — and thus an orientation — of the plane in accordance with the right-hand rule.



**FIG. 2.11**

Vector cross product in three dimensions.

1. is distributive over addition,
2. permits commutative multiplication by scalars, and
3. satisfies  $a^2 = |a|^2$  for an arbitrary vector  $a$

then it will be possible to define the inverse vector  $a^{-1}$  of a nonzero vector  $a$ . Thus one way to develop the desired product of vectors is to use these notions and formulate an axiomatic approach to multiplying vectors. This will have the advantage of not being limited to a particular number of dimensions.

In particular, assume an  $N$ -dimensional vector space  $\mathbb{R}^N$  with vectors  $a, b, c, \dots$ . The *geometric algebra* associated with this vector space will be denoted  $\mathcal{G}(\mathbb{R}^N)$ . It is closed under addition and multiplication, i.e., addition or multiplication of two members of  $\mathcal{G}(\mathbb{R}^N)$  results in a member of  $\mathcal{G}(\mathbb{R}^N)$ . Motivated by the lessons learned before with the  $\odot$  product as well as the complexified cross product approach, one may now adopt the following axioms for members  $A, B$ —called *multivectors*—of the geometric algebra  $\mathcal{G}(\mathbb{R}^N)$ :

1. Addition is commutative:

$$A + B = B + A \quad (2.34)$$

2. Addition and multiplication are associative:

$$\begin{aligned} (A + B) + C &= A + (B + C) \\ (AB)C &= A(BC) \end{aligned} \quad (2.35)$$

3. Multiplication is distributive with respect to addition:

$$\begin{aligned} A(B + C) &= AB + AC \\ (B + C)A &= BC + CA \end{aligned} \quad (2.36)$$

It is important to note that except for multiplication by a scalar, multiplication is not necessarily commutative. Hence distributivity is defined from both sides.

4. There exist unique additive and multiplicative identities:

$$\begin{aligned} A + 0 &= A \\ 1A &= A \end{aligned} \tag{2.37}$$

5. Every multivector has a unique additive inverse:

$$A + (-A) = 0 \tag{2.38}$$

However, every multivector does not necessarily have a *multiplicative* inverse.

6. The square of any nonzero vector  $a$  is equal to a positive<sup>2</sup> scalar:

$$aa = |a|^2 > 0 \tag{2.39}$$

This axiom is called the *contraction rule*.

Other than sacrificing commutativity of multiplication, these axioms are consistent with the algebra of real numbers. In effect, multivectors are treated as much like real numbers as possible. As a result, it follows from the axioms that the real numbers  $\mathbb{R}$  and the vectors in  $\mathbb{R}^N$  are themselves members of  $\mathcal{G}(\mathbb{R}^N)$ .

To begin to understand the implications of these axioms, apply them to the multiplication of two vectors. In particular, let  $a$  and  $b$  be two  $N$ -dimensional vectors and examine the *geometric product* denoted:

$$ab \tag{2.40}$$

Because  $\mathcal{G}(\mathbb{R}^N)$  is closed under multiplication (i.e., the geometric product), the multivectors  $ab$  and  $ba$  are both members of  $\mathcal{G}(\mathbb{R}^N)$ . Because the axioms permit addition of multivectors and multiplication by scalars, one may now write

$$ab = \frac{1}{2}(ab + ba) + \frac{1}{2}(ab - ba) \tag{2.41}$$

The first term  $\frac{1}{2}(ab + ba)$  is symmetric, i.e., interchanging  $a$  and  $b$  does not change the result. By contrast, the second term  $\frac{1}{2}(ab - ba)$  is antisymmetric—interchanging  $a$  and  $b$  changes the sign of the result. In this regard, the first term shares symmetry with the inner product, whereas the second term shares antisymmetry with the cross product. Further, the second term is also like the cross product in that it vanishes if  $a=b$  (i.e.,  $a \times a = 0$ ).

Now examine the geometric product of  $a - b$  with itself

$$(a - b)(a - b) \tag{2.42}$$

---

<sup>2</sup>The requirement that the scalar be *positive* can be relaxed to allow for application of geometric algebra in other areas, for example, electromagnetics in the context of relativity.

Using distributivity leads to

$$(\mathbf{a} - \mathbf{b})(\mathbf{a} - \mathbf{b}) = \mathbf{aa} - \mathbf{ba} - \mathbf{ab} + \mathbf{bb} \quad (2.43)$$

Applying the contraction rule to both sides of this relation yields

$$|\mathbf{a} - \mathbf{b}|^2 = |\mathbf{a}|^2 - (\mathbf{ab} + \mathbf{ba}) + |\mathbf{b}|^2 \quad (2.44)$$

Because the left-hand side is a scalar, the right-hand side must also be a scalar. It follows therefore that  $\mathbf{ab} + \mathbf{ba}$  is a scalar. With this in mind one may now define the *inner product* (sometimes called the *dot product* or the *scalar product*) to be

$$\mathbf{a} \cdot \mathbf{b} = \frac{1}{2}(\mathbf{ab} + \mathbf{ba}) \quad (2.45)$$

Note that the symmetry of the right-hand side yields

$$\mathbf{a} \cdot \mathbf{b} = \mathbf{b} \cdot \mathbf{a} \quad (2.46)$$

i.e., the inner product is commutative, as expected.

With this definition the geometric product becomes

$$\mathbf{ab} = \mathbf{a} \cdot \mathbf{b} + \frac{1}{2}(\mathbf{ab} - \mathbf{ba}) \quad (2.47)$$

This result is starting to look like the  $\odot$  product before, except that the last term on the right-hand side should not be identified with the cross product (either directly or complexified) because the cross product is not available in all dimensions.

To continue, now define an associative antisymmetric product of two vectors as

$$\mathbf{a} \wedge \mathbf{b} = \frac{1}{2}(\mathbf{ab} - \mathbf{ba}) \quad (2.48)$$

This product is called the *outer product* (also known as the *wedge product* or the *exterior product*) and was introduced by Grassmann in 1844. From this definition one can immediately deduce two important properties:

$$\mathbf{a} \wedge \mathbf{b} = -\mathbf{b} \wedge \mathbf{a} \quad (2.49)$$

$$\mathbf{a} \wedge \mathbf{a} = 0 \quad (2.50)$$

Because the axioms require the geometric product to be associative, it follows that the outer product must also be associative:

$$(\mathbf{a} \wedge \mathbf{b}) \wedge \mathbf{c} = \mathbf{a} \wedge (\mathbf{b} \wedge \mathbf{c}) \quad (2.51)$$

As of yet it is not clear what kind of object  $\mathbf{a} \wedge \mathbf{b}$  is, and this topic will be explored further later. For now, however, examine its consequences for the geometric product, which now takes the form:

$$\mathbf{ab} = \mathbf{a} \cdot \mathbf{b} + \mathbf{a} \wedge \mathbf{b} \quad (2.52)$$

This product was introduced by Clifford in 1878.

From the properties of the inner and outer products it follows that

$$\mathbf{ba} = \mathbf{b} \cdot \mathbf{a} + \mathbf{b} \wedge \mathbf{a} = \mathbf{a} \cdot \mathbf{b} - \mathbf{a} \wedge \mathbf{b} \quad (2.53)$$

Therefore in general

$$\mathbf{ab} \neq \mathbf{ba} \quad (2.54)$$

The geometric product of two vectors is *not necessarily commutative* except in special cases. For example, examine the special case where  $\mathbf{a}$  is orthogonal to  $\mathbf{b}$ , i.e.,  $\mathbf{a} \cdot \mathbf{b} = 0$ , which may now be taken as the definition of orthogonality of vectors. Then

$$\mathbf{ab} = \mathbf{a} \wedge \mathbf{b} = -\mathbf{b} \wedge \mathbf{a} = -\mathbf{ba} \quad (2.55)$$

*The geometric product of orthogonal vectors is antisymmetric and equal to their outer product.* Similarly, examine the special case where  $\mathbf{a}$  is parallel to  $\mathbf{b}$ —i.e.,  $\mathbf{a} \wedge \mathbf{b} = 0$ , which may now be taken as the definition of parallel vectors (e.g., let  $\mathbf{b} = \lambda \mathbf{a}$  for some scalar  $\lambda$ —then  $\mathbf{a} \wedge \mathbf{b} = \mathbf{a} \wedge \lambda \mathbf{a} = \lambda \mathbf{a} \wedge \mathbf{a} = 0$ ). Then

$$\mathbf{ab} = \mathbf{a} \cdot \mathbf{b} = \mathbf{b} \cdot \mathbf{a} = \mathbf{ba} \quad (2.56)$$

*The geometric product of parallel vectors is symmetric and equal to their inner product.* Thus the symmetry or antisymmetry of the geometric product is a measure of whether vectors are parallel or orthogonal. In general, two vectors are neither parallel nor orthogonal, and their geometric product is neither symmetric nor antisymmetric.

Apply the contraction rule to an arbitrary vector  $\mathbf{a}$

$$\mathbf{a}^2 = |\mathbf{a}|^2 \quad (2.57)$$

Now define the multiplicative inverse vector of a nonzero vector  $\mathbf{a}$  as was done previously:

$$\mathbf{a}^{-1} \triangleq \frac{\mathbf{a}}{|\mathbf{a}|^2} \quad (2.58)$$

Hence

$$\mathbf{a}^{-1}\mathbf{a} = \mathbf{aa}^{-1} = \frac{\mathbf{aa}}{|\mathbf{a}|^2} = 1 \quad (2.59)$$

Using associativity, it follows immediately that

$$\mathbf{a}^{-1}\mathbf{ab} = \mathbf{b} \quad (2.60)$$

Similarly we have

$$\mathbf{abb}^{-1} = \mathbf{a} \quad (2.61)$$

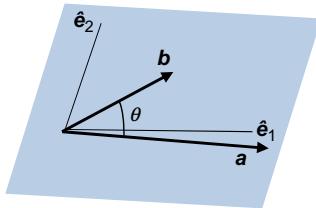
This shows that one can perform division with nonzero vectors through left- or right-multiplication by inverse vectors as appropriate.

Thus far the discussion has been somewhat abstract. To make this development somewhat more concrete for engineers, consider two nonparallel vectors expressed in a basis in their shared plane as shown in Fig. 2.12.

In particular, let  $\{\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2\}$  be an orthonormal basis for a Euclidean plane and represent two vectors  $\mathbf{a}$  and  $\mathbf{b}$  in this plane as

$$\mathbf{a} = a_1 \hat{\mathbf{e}}_1 + a_2 \hat{\mathbf{e}}_2 \quad (2.62)$$

$$\mathbf{b} = b_1 \hat{\mathbf{e}}_1 + b_2 \hat{\mathbf{e}}_2 \quad (2.63)$$

**FIG. 2.12**

Two vectors define a plane.

Note that because the two nonparallel vectors themselves define the plane, this plane and these vectors can be in an  $N$ -dimensional space where  $N > 2$ . Then

$$\begin{aligned} \mathbf{ab} &= (a_1 \hat{\mathbf{e}}_1 + a_2 \hat{\mathbf{e}}_2)(b_1 \hat{\mathbf{e}}_1 + b_2 \hat{\mathbf{e}}_2) \\ &= a_1 b_1 \hat{\mathbf{e}}_1 \hat{\mathbf{e}}_1 + a_2 b_2 \hat{\mathbf{e}}_2 \hat{\mathbf{e}}_2 + a_1 b_2 \hat{\mathbf{e}}_1 \hat{\mathbf{e}}_2 + a_2 b_1 \hat{\mathbf{e}}_2 \hat{\mathbf{e}}_1 \end{aligned} \quad (2.64)$$

Note that axioms have been used to remove the parentheses, and the terms on the right-hand side have been rearranged. From the contraction rule and the fact that the vectors in this expression are orthonormal (in which case  $\hat{\mathbf{e}}_1 \cdot \hat{\mathbf{e}}_2 = 0$ ), one finds

$$\hat{\mathbf{e}}_1 \hat{\mathbf{e}}_1 = \hat{\mathbf{e}}_2 \hat{\mathbf{e}}_2 = 1 \quad (2.65)$$

$$\hat{\mathbf{e}}_1 \hat{\mathbf{e}}_2 = \hat{\mathbf{e}}_1 \wedge \hat{\mathbf{e}}_2 = -\hat{\mathbf{e}}_2 \wedge \hat{\mathbf{e}}_1 = -\hat{\mathbf{e}}_2 \hat{\mathbf{e}}_1 \quad (2.66)$$

From these relations it then follows

$$\mathbf{ab} = a_1 b_1 + a_2 b_2 + (a_1 b_2 - a_2 b_1) \hat{\mathbf{e}}_1 \hat{\mathbf{e}}_2 \quad (2.67)$$

The reader will recognize the term

$$\mathbf{a} \cdot \mathbf{b} = a_1 b_1 + a_2 b_2 \quad (2.68)$$

as the inner product computed from the components of the two vectors. It follows from the definition of the geometric product that the other term is given by

$$\mathbf{a} \wedge \mathbf{b} = (a_1 b_2 - a_2 b_1) \hat{\mathbf{e}}_1 \hat{\mathbf{e}}_2 \quad (2.69)$$

These expressions for the inner and outer products allow one to *compute* these quantities in terms of a basis. However, as already shown before, adopting a basis is not necessary to *define* these concepts, which are fundamentally geometric ideas.

To explore this latter term, temporarily restrict attention to a three-dimensional space that has the right-handed basis  $\{\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \hat{\mathbf{e}}_3\}$ . In this case the cross product of  $\mathbf{a}$  and  $\mathbf{b}$  is given by

$$\mathbf{a} \times \mathbf{b} = (a_1 \hat{\mathbf{e}}_1 + a_2 \hat{\mathbf{e}}_2) \times (b_1 \hat{\mathbf{e}}_1 + b_2 \hat{\mathbf{e}}_2) = (a_1 b_2 - a_2 b_1) \hat{\mathbf{e}}_3 \quad (2.70)$$

The basis vector  $\hat{\mathbf{e}}_3$  determines the normal to the plane defined by  $\hat{\mathbf{e}}_1$  and  $\hat{\mathbf{e}}_2$  in accordance with a right-handed coordinate system. Note that when  $a_1 b_2 - a_2 b_1 > 0$ , then  $\hat{\mathbf{n}} = \hat{\mathbf{e}}_3$ . Otherwise  $\hat{\mathbf{n}} = -\hat{\mathbf{e}}_3$ . The magnitude of the cross product vector is

$$|a_1 b_2 - a_2 b_1| = |\mathbf{a}| |\mathbf{b}| \sin \theta \quad (2.71)$$

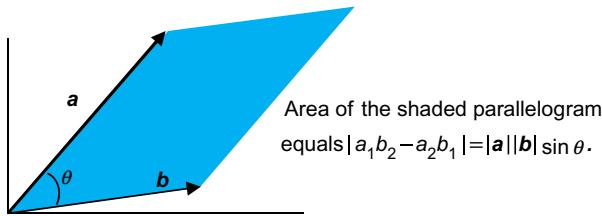


FIG. 2.13

Magnitude of a cross product.

This relation defines the area of the parallelogram formed by the two vectors as illustrated in Fig. 2.13. But this same quantity also appears as part of the outer product.

Therefore one may write

$$\mathbf{a} \wedge \mathbf{b} = |\mathbf{a}| |\mathbf{b}| \sin \theta (\pm \hat{\mathbf{e}}_1 \hat{\mathbf{e}}_2) \quad (2.72)$$

where  $0 \leq \theta \leq \pi$  and the sign is given by the sign of  $(a_1 b_2 - a_2 b_1)$ . Call

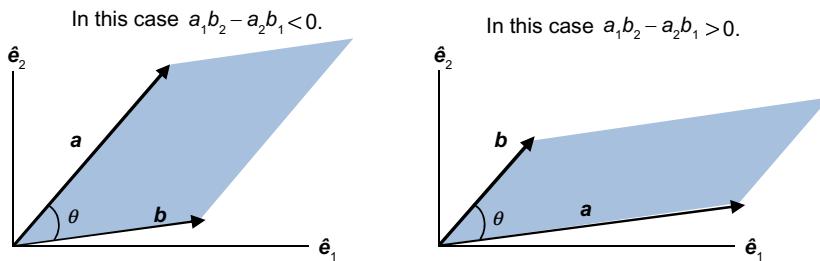
$$|\mathbf{a} \wedge \mathbf{b}| = |\mathbf{a}| |\mathbf{b}| \sin \theta \quad (2.73)$$

the *magnitude* of  $\mathbf{a} \wedge \mathbf{b}$ . The particular choice of sign is determined by the relative orientation of the two vectors as illustrated in Fig. 2.14:

From Fig. 2.14 one can see that the sign of  $(a_1 b_2 - a_2 b_1)$  supplies the outer product with an *orientation*. This follows from the fact that  $\hat{\mathbf{e}}_1 \hat{\mathbf{e}}_2 = -\hat{\mathbf{e}}_2 \hat{\mathbf{e}}_1$ . Thus we have

$$\mathbf{a} \wedge \mathbf{b} = |\mathbf{a}| |\mathbf{b}| \sin \theta \begin{cases} \hat{\mathbf{e}}_1 \hat{\mathbf{e}}_2, a_1 b_2 - a_2 b_1 > 0 \\ \hat{\mathbf{e}}_2 \hat{\mathbf{e}}_1, a_1 b_2 - a_2 b_1 < 0 \end{cases} \quad (2.74)$$

The cross product of two vectors comprises both a magnitude and a vector—the magnitude is determined by the geometric relationship of the vectors *in the plane* whereas the unit vector is *normal to the plane*. The difficulty with the cross product for purposes of defining an invertible, associative product of vectors in any



Area of the parallelogram is the same in both cases.

FIG. 2.14

Area defined by outer product is independent of orientation.

dimension is that the vector defined by the cross product *leaves the plane*. It is a consequence of leaving the plane that restricts the cross product to three dimensions, since only in three dimensions can there be a uniquely defined normal to a plane.

The outer product, on the other hand, comprises the same magnitude as the cross product but replaces the normal vector, which leaves the plane, with an object (no longer a vector) that *remains in the plane*. As a result, the outer product can be defined in any number of dimensions because two vectors always define a plane regardless of the dimension of the space (unless they are parallel, in which case the quantity being considered equals zero anyway).

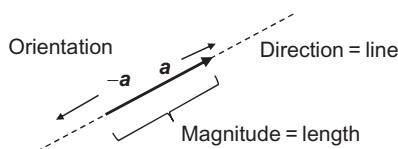
The outer product  $\mathbf{a} \wedge \mathbf{b}$  defines an *oriented planar object* called a *bivector*. In this regard, the bivector  $\mathbf{a} \wedge \mathbf{b}$  is an extension of the vector idea to a plane. As was discussed earlier, a vector has three attributes as shown in Fig. 2.15:

1. A magnitude, i.e., its length;
2. A direction, i.e., the line along which it is directed;
3. An orientation, i.e., the sign of the vector.

Similarly, a *bivector*, has three attributes as shown in Fig. 2.16:

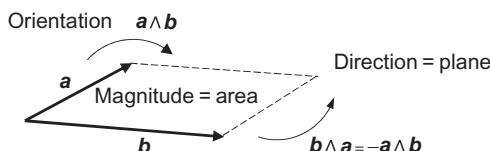
1. A magnitude, i.e., the area of the parallelogram;
2. A direction, i.e., the plane in which it is defined;
3. An orientation, i.e., the sign of the bivector.

The orientation—and hence the sign—is provided by the order of the vectors in the outer product. One obtains  $\mathbf{a} \wedge \mathbf{b}$  by first traversing  $\mathbf{a}$  and then  $\mathbf{b}$ . Similarly one obtains  $\mathbf{b} \wedge \mathbf{a}$  by traversing in the opposite direction. Both give rise to the same area in the same plane—only the orientation as embodied by the sign is different.



**FIG. 2.15**

Attributes of a vector.



**FIG. 2.16**

Attributes of a bivector.

Now examine

$$\hat{\mathbf{e}}_1 \hat{\mathbf{e}}_2 = \hat{\mathbf{e}}_1 \wedge \hat{\mathbf{e}}_2 \quad (2.75)$$

From the discussion before, the magnitude of both  $\hat{\mathbf{e}}_1 \wedge \hat{\mathbf{e}}_2$  and  $\hat{\mathbf{e}}_2 \wedge \hat{\mathbf{e}}_1$  is given by

$$|\hat{\mathbf{e}}_1| |\hat{\mathbf{e}}_2| \sin \frac{\pi}{2} = 1 \quad (2.76)$$

Thus  $\hat{\mathbf{e}}_1 \wedge \hat{\mathbf{e}}_2$  and  $\hat{\mathbf{e}}_2 \wedge \hat{\mathbf{e}}_1$  are “unit” bivectors associated with the plane defined by the vectors  $\hat{\mathbf{e}}_1$  and  $\hat{\mathbf{e}}_2$ . In addition, they are also *oriented* in the sense that they give rise to two possible directions:

$$\hat{\mathbf{e}}_1 \wedge \hat{\mathbf{e}}_2 = -\hat{\mathbf{e}}_2 \wedge \hat{\mathbf{e}}_1 \quad (2.77)$$

Observe now that the same unit bivectors arise *regardless of how the basis vectors are chosen* in the plane. In particular,

- (1) The magnitude is always 1;
- (2) Any two orthonormal vectors in a plane define the same plane; and
- (3) The orientation can always be specified by choosing the sign.

Instead of using  $\pm \hat{\mathbf{e}}_1 \hat{\mathbf{e}}_2$ , which depends on the basis, designate the unit bivector in the plane defined by the vectors  $\mathbf{a}$  and  $\mathbf{b}$  as  $I_{ab}$ , which is independent of the particular basis. Thus  $\mathbf{a} \wedge \mathbf{b}$  may be written as

$$\mathbf{a} \wedge \mathbf{b} = |\mathbf{a}| |\mathbf{b}| \sin \theta I_{ab} \quad (2.78)$$

[Fig. 2.16](#) shows that  $I_{ab}$  automatically supplies the same orientation as  $\pm \hat{\mathbf{e}}_1 \hat{\mathbf{e}}_2$  in their respective cases. With this result, the geometric product becomes

$$\mathbf{a} \mathbf{b} = |\mathbf{a}| |\mathbf{b}| (\cos \theta + \sin \theta I_{ab}) \quad (2.79)$$

This is a completely *coordinate-free* treatment of the product of two vectors in *any dimension*. The motivation for the name “geometric product” is now clear—the geometric product of two vectors depends only on the magnitudes of the vectors and the smaller angle between them in their shared plane, both of which are geometric properties. It does **not** require the specification of any basis.

## 2.2.2 GEOMETRIC ALGEBRA

### 2.2.2.1 Geometric Algebra in Two Dimensions

In two dimensions, the geometric product leads to bivectors in addition to the scalars and vectors that already were in the underlying vector space  $\mathbb{R}^2$ . Does the geometric product of these quantities with each other lead to any other new objects? To answer this question one must explore the properties of  $I_{ab}$  further. As a matter of notation, designate the unit bivector in an unspecified plane as  $I_2$  and reserve the notation  $I_{ab}$  for the unit bivector in the specific plane defined by vectors  $\mathbf{a}$  and  $\mathbf{b}$  and oriented the same as  $\mathbf{a} \wedge \mathbf{b}$ . First consider the quantity  $I_2^2$ . Choose an arbitrary orthonormal basis  $\{\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2\}$  in the plane and write

$$I_2 = \hat{\mathbf{e}}_1 \wedge \hat{\mathbf{e}}_2 = \hat{\mathbf{e}}_1 \hat{\mathbf{e}}_2 \quad (2.80)$$

The square of this quantity is given by

$$I_2^2 = (\hat{e}_1 \hat{e}_2)(\hat{e}_1 \hat{e}_2) \quad (2.81)$$

Because the geometric product is associative the parentheses are unnecessary:

$$I_2^2 = \hat{e}_1 \hat{e}_2 \hat{e}_1 \hat{e}_2 \quad (2.82)$$

Recall now that the geometric product of orthogonal vectors is antisymmetric. Therefore  $\hat{e}_1 \hat{e}_2 = -\hat{e}_2 \hat{e}_1$  and the terms may be reordered to obtain

$$I_2^2 = -\hat{e}_1 \hat{e}_1 \hat{e}_2 \hat{e}_2 \quad (2.83)$$

Application of the contraction rule twice now yields

$$I_2^2 = -1 \quad (2.84)$$

This calculation is independent of the actual orthonormal basis and shows that a unit bivector is a *geometric* object with the same algebraic property as  $i = \sqrt{-1}$ ! This result has two immediate consequences—it gives rise to rotations, and it closes the algebra.

First to see how it gives rise to rotations, examine

$$I_2 \hat{e}_1 = \hat{e}_1 \hat{e}_2 \hat{e}_1 = -\hat{e}_1 \hat{e}_1 \hat{e}_2 = -\hat{e}_2 \quad (2.85)$$

$$I_2 \hat{e}_2 = \hat{e}_1 \hat{e}_2 \hat{e}_2 = \hat{e}_1 \quad (2.86)$$

This shows that multiplication of a vector by  $I_2$  from the left causes a rotation of that vector by 90 degrees. This property is illustrated in Fig. 2.17.

Similarly

$$\hat{e}_1 I_2 = \hat{e}_1 \hat{e}_1 \hat{e}_2 = \hat{e}_2 \quad (2.87)$$

$$\hat{e}_2 I_2 = \hat{e}_2 \hat{e}_1 \hat{e}_2 = -\hat{e}_1 \quad (2.88)$$

Multiplication of a vector by  $I_2$  from the right causes a rotation of the vector by 90 degrees in the opposite direction. These are properties that arise in the complex plane with multiplication of complex numbers by  $-i$  and  $i$ .

Second, examine products of the various kinds of objects, namely, scalars, vectors, and bivectors. As is well known, the product of a scalar with a scalar is another scalar, and the product of a scalar with a vector is a vector. It is straightforward to see from the definition of the outer product that the product of a scalar with a bivector is a bivector. Thus multiplying any of the three objects—scalar, vector, or bivector—by a

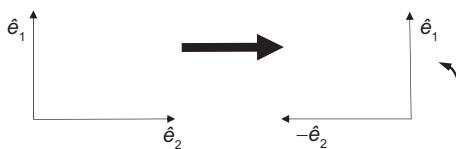


FIG. 2.17

Multiplication by a unit bivector induces rotations.

scalar leads to that same kind of object. Now consider multiplication by vectors. As already discussed, multiplication of a vector by a scalar yields a vector, and multiplication of two vectors yields the sum of a scalar and a bivector. Since such sums are included in  $\mathcal{G}(\mathbb{R}^2)$ , the product of two vectors yields an object in  $\mathcal{G}(\mathbb{R}^2)$ . Now consider the product of a vector and a bivector. First, in two dimensions, there is only one plane. Therefore it follows that any bivector  $B$  may be written as

$$B = \pm|B|\hat{\mathbf{e}}_1\hat{\mathbf{e}}_2 = \pm|B|I_2 \quad (2.89)$$

where  $\{\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2\}$  defines an arbitrary orthonormal basis in the plane. Similarly, any vector  $\mathbf{a}$  may be written as

$$\mathbf{a} = a_1\hat{\mathbf{e}}_1 + a_2\hat{\mathbf{e}}_2 \quad (2.90)$$

Examine now the products  $\mathbf{a}B$  and  $B\mathbf{a}$ :

$$\begin{aligned} \mathbf{a}B &= \pm a_1|B|\hat{\mathbf{e}}_1\hat{\mathbf{e}}_1\hat{\mathbf{e}}_2 \pm a_2|B|\hat{\mathbf{e}}_2\hat{\mathbf{e}}_1\hat{\mathbf{e}}_2 \\ &= \pm a_1|B|\hat{\mathbf{e}}_2 \mp a_2|B|\hat{\mathbf{e}}_1 \end{aligned} \quad (2.91)$$

$$\begin{aligned} B\mathbf{a} &= \pm a_1|B|\hat{\mathbf{e}}_1\hat{\mathbf{e}}_2\hat{\mathbf{e}}_1 \pm a_2|B|\hat{\mathbf{e}}_1\hat{\mathbf{e}}_2\hat{\mathbf{e}}_2 \\ &= \mp a_1|B|\hat{\mathbf{e}}_2 \pm a_2|B|\hat{\mathbf{e}}_1 \end{aligned} \quad (2.92)$$

These quantities are vectors. Thus multiplication of a vector by a bivector yields a vector. From this result it follows that multiplication of any object—scalar, vector, or bivector—by a vector results in another object in  $\mathcal{G}(\mathbb{R}^2)$ .

Finally, examine multiplication by a bivector. Multiplication of a scalar by a bivector leads to a bivector, and multiplication of a vector by a bivector yields a vector. Therefore the only multiplication not yet considered is the multiplication of two bivectors:

$$B_1B_2 = \pm|B_1||B_2|I_2^2 \quad (2.93)$$

But has already been shown before,  $I_2^2 = -1$ . Therefore

$$B_1B_2 = \mp|B_1||B_2| \quad (2.94)$$

This is a scalar, which is in  $\mathcal{G}(\mathbb{R}^2)$ . As a result, *the algebra is closed* under the geometric product—multiplication of any two objects in  $\mathcal{G}(\mathbb{R}^2)$  results in another object in  $\mathcal{G}(\mathbb{R}^2)$  where in general a multivector in  $\mathcal{G}(\mathbb{R}^2)$  takes the form

$$A = a_0 + \mathbf{a} + a_3I_2 \quad (2.95)$$

From the general form in Eq. (2.95) it may be seen that  $\mathcal{G}(\mathbb{R}^2)$  is a four-dimensional linear space. That is, the quantities  $1, \hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, I_2$  form a basis for expressing a general multivector in  $\mathcal{G}(\mathbb{R}^2)$ .

If one now thinks of  $a_0 + a_3I_2$  as a “complex number” (discussed further later) then the general object in  $\mathcal{G}(\mathbb{R}^2)$  takes the form of the sum of a vector and a complex number. Furthermore, since the product of a scalar or a bivector with a vector leads in both cases to another vector, it follows that

$$(a_0 + a_3I_2)\mathbf{a} = \mathbf{b} \quad (2.96)$$

i.e., the product of the “complex number”  $a_0 + a_3 I_2$  with a vector results in another vector. Since any vector in the plane can be obtained from any other vector in the plane by a rotation and a dilation, one may think of  $a_0 + a_3 I_2$  as an operator that rotates and dilates vectors into other vectors. Such a quantity is called a *spinor*. Spinors are discussed further later.

### 2.2.2.2 Geometric Algebra in Three Dimensions

Consider now a three-dimensional space with an arbitrary orthonormal basis  $\{\hat{e}_1, \hat{e}_2, \hat{e}_3\}$  and define

$$I_3 = \hat{e}_1 \wedge \hat{e}_2 \wedge \hat{e}_3 = \hat{e}_1 \hat{e}_2 \hat{e}_3 \quad (2.97)$$

This quantity defines an object in the algebra called a unit trivector. Just as a vector is associated with a line and a bivector is associated with a plane, a trivector is an object associated with a volume, as shown in Fig. 2.18.

Interchanging any two of the basis vectors in this unit trivector changes the sign (from the antisymmetry of the outer product). Thus a trivector is an *oriented magnitude associated with a volume*. Like the vector and the bivector, it has three attributes:

1. A magnitude—a volume;
2. A direction—the three-dimensional space where it is defined; and
3. An orientation—the sign of the trivector.

Now examine the product

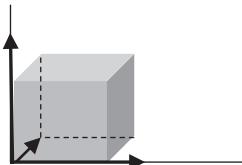
$$I_3 \hat{e}_3 = \hat{e}_1 \hat{e}_2 \hat{e}_3 \hat{e}_3 \quad (2.98)$$

From the contraction rule and the fact that  $\hat{e}_3$  is a unit vector it follows immediately that

$$I_3 \hat{e}_3 = \hat{e}_1 \hat{e}_2 \quad (2.99)$$

Using the antisymmetry of the geometric product of orthogonal vectors one also finds

$$\begin{aligned} \hat{e}_3 I_3 &= \hat{e}_3 \hat{e}_1 \hat{e}_2 \hat{e}_3 \\ &= -\hat{e}_1 \hat{e}_3 \hat{e}_2 \hat{e}_3 \\ &= \hat{e}_1 \hat{e}_2 \hat{e}_3 \hat{e}_3 \\ &= I_3 \hat{e}_3 \end{aligned} \quad (2.100)$$



**FIG. 2.18**

Trivector as outer product of three vectors.

With these relations now return to the result for the cross product  $\mathbf{a} \times \mathbf{b}$  and examine

$$I_3(\mathbf{a} \times \mathbf{b}) = I_3|\mathbf{a}||\mathbf{b}|\sin\theta \hat{\mathbf{e}}_3 \quad (2.101)$$

Recall that scalars commute with all objects in the algebra. Thus

$$I_3(\mathbf{a} \times \mathbf{b}) = |\mathbf{a}||\mathbf{b}|\sin\theta I_3 \hat{\mathbf{e}}_3 \quad (2.102)$$

Using the result for  $I_3 \hat{\mathbf{e}}_3$  previously, we now have

$$\begin{aligned} I_3(\mathbf{a} \times \mathbf{b}) &= |\mathbf{a}||\mathbf{b}|\sin\theta \hat{\mathbf{e}}_1 \hat{\mathbf{e}}_2 \\ &= \mathbf{a} \wedge \mathbf{b} \end{aligned} \quad (2.103)$$

From  $I_3 \hat{\mathbf{e}}_3 = \hat{\mathbf{e}}_3 I_3$  one also finds

$$(\mathbf{a} \times \mathbf{b}) I_3 = \mathbf{a} \wedge \mathbf{b} \quad (2.104)$$

This shows that the cross product and the outer product are related through multiplication by  $I_3$ .

Now, by sifting  $\hat{\mathbf{e}}_3$  and then  $\hat{\mathbf{e}}_2$  back through  $I_3 = \hat{\mathbf{e}}_1 \hat{\mathbf{e}}_2 \hat{\mathbf{e}}_3$  and changing signs with each interchange of orthogonal vectors, it is straightforward to show  $\hat{\mathbf{e}}_3 \hat{\mathbf{e}}_2 \hat{\mathbf{e}}_1 = -\hat{\mathbf{e}}_1 \hat{\mathbf{e}}_2 \hat{\mathbf{e}}_3 = -I_3$ . From this one obtains

$$-I_3 I_3 = \hat{\mathbf{e}}_3 \hat{\mathbf{e}}_2 \hat{\mathbf{e}}_1 \hat{\mathbf{e}}_2 \hat{\mathbf{e}}_3 = 1 \quad (2.105)$$

Hence the unit trivector  $I_3$  has an inverse

$$I_3^{-1} = \hat{\mathbf{e}}_3 \hat{\mathbf{e}}_2 \hat{\mathbf{e}}_1 = -I_3 \quad (2.106)$$

Therefore one may also write

$$\mathbf{a} \wedge \mathbf{b} I_3^{-1} = \mathbf{a} \times \mathbf{b} I_3 I_3^{-1} = \mathbf{a} \times \mathbf{b} \quad (2.107)$$

This relation defines the cross product (a vector) as the three-dimensional *dual*<sup>3</sup> of the outer product (a bivector). As shown in Fig. 2.19, in three dimensions the duality operation interchanges a plane with a vector that is orthogonal to that plane.

The more general notion of duality in  $N$  dimensions is discussed later.

From the previous result it follows that in three dimensions

$$I_3^2 = -1 \quad (2.108)$$

Hence more than one geometric quantity squares to  $-1$ ! Using the previous duality relationship one may now write in three dimensions

$$\mathbf{ab} = \mathbf{a} \cdot \mathbf{b} + I_3 \mathbf{a} \times \mathbf{b} \quad (2.109)$$

This is similar to both the  $\odot$  product and the “complex cross product” approach explored before, except that the cross product  $\mathbf{a} \times \mathbf{b}$  or  $i\mathbf{a} \times \mathbf{b}$  is replaced by  $I_3 \mathbf{a} \times \mathbf{b}$ . Although  $i$  shares the pertinent algebraic property with both  $I_2$  and  $I_3$ —namely, they each square to  $-1$ —the quantities  $I_2$  and  $I_3$  are quite different objects both geometrically and algebraically. Moreover, in three dimensions, each plane has its own unit bivector. Thus even in three dimensions there are a plethora of quantities that square

---

<sup>3</sup>Some authors call this the Hodge dual.

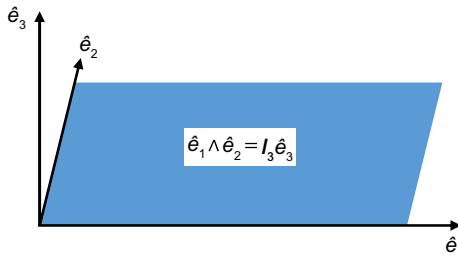


FIG. 2.19

Cross product is dual to outer product.

to  $-1$ . As a result, it is not clear *a priori* which object that squares to  $-1$  should be used to define the associative product of vectors, and use of the unit imaginary number  $i$  in this regard would be arbitrary. Moreover, use of the unit imaginary  $i$  would sacrifice much of the geometric interpretation that we obtained from the use of the multivectors in  $\mathcal{G}(\mathbb{R}^N)$ .

The general multivector in three dimensions takes the form

$$A = \alpha_0 + \alpha_1 \hat{e}_1 + \alpha_2 \hat{e}_2 + \alpha_3 \hat{e}_3 + \alpha_{12} \hat{e}_1 \hat{e}_2 + \alpha_{23} \hat{e}_2 \hat{e}_3 + \alpha_{13} \hat{e}_1 \hat{e}_3 + \alpha_{123} I_3 \quad (2.110)$$

Observe that Eq. (2.110) shows that  $\mathcal{G}(\mathbb{R}^3)$  is an eight-dimensional linear space. In other words, the quantities  $1, \hat{e}_1, \hat{e}_2, \hat{e}_3, \hat{e}_1 \hat{e}_2, \hat{e}_2 \hat{e}_3, \hat{e}_3 \hat{e}_1, I_3$  form a basis for expressing a general multivector in  $\mathcal{G}(\mathbb{R}^3)$ . Each vector  $\hat{e}_1, \hat{e}_2, \hat{e}_3$  forms a one-dimensional subspace of  $\mathbb{R}^3$ , of which there are three such subspaces. Moreover, there are three ways to choose these vectors two at a time and each such choice defines a two-dimensional subspace of  $\mathbb{R}^3$ . Finally, there is only one way to choose the vectors three at a time.

### 2.2.2.3 Geometric Algebra in Three Dimensions

Consider now  $\mathbb{R}^N$  with orthonormal basis vectors  $\hat{e}_1, \hat{e}_2, \dots, \hat{e}_N$ . By choosing these basis vectors  $k$  at time where  $k = 1, 2, \dots, N$ , one can define  $\binom{N}{k}$   $k$ -dimensional subspaces. With each such subspace one may associate the unit  $k$ -vector formed from the pertinent basis vectors. In this fashion, one can form the general multivector in  $\mathcal{G}(\mathbb{R}^N)$  as

$$A = \alpha_0 + \alpha_1 \hat{e}_1 + \dots + \alpha_2 \hat{e}_N + \alpha_{12} \hat{e}_1 \hat{e}_2 + \dots + \alpha_{(N-1)N} \hat{e}_{N-1} \hat{e}_N + \alpha_{123} \hat{e}_1 \hat{e}_2 \hat{e}_3 + \dots + \alpha_{(N-2)(N-1)N} \hat{e}_{N-2} \hat{e}_{N-1} \hat{e}_N + \dots + \alpha_{1\dots N} I_N \quad (2.111)$$

The number of unit basis elements obtained in this way is given by

$$\sum_{k=0}^N \binom{N}{k} = 2^N$$

Thus  $\mathcal{G}(\mathbb{R}^N)$  is a  $2N$ -dimensional linear space. Because there are  $\binom{N}{k}$  unit  $k$ -blades as part of the set of basis elements, it follows that the basis set comprises 1 scalar basis element,  $N$  basis vectors,  $\frac{N(N-1)}{2}$  basis bivectors, etc. In particular, there is only one basis  $N$ -blade. The unit  $N$ -blade  $I_N$  is called a *pseudoscalar*. In two dimensions, the unit pseudoscalar is  $I_2$ , which is a bivector. In three dimensions, the unit pseudoscalar is  $I_3$ , which is a trivector, and so on. As was shown before,  $I_2^2 = -1$  and  $I_3^2 = -1$ . Therefore in those dimensions, the product of a pseudoscalar with another pseudoscalar is a scalar. This property actually holds in general. One consequence is that the algebra  $\mathcal{G}(\mathbb{R}^N)$  is closed under the geometric product. This means that the geometric product of any two elements of  $\mathcal{G}(\mathbb{R}^N)$  yields another member of  $\mathcal{G}(\mathbb{R}^N)$ . To understand how this occurs, we must understand how the geometric product applies to general multivectors rather than just to vectors as has been the case thus far. This will be the focus of the next section. Prior to examining the general product of multivectors, however, we examine some other consequences of the geometric product of vectors.

First define for an arbitrary multivector in  $\mathcal{G}(\mathbb{R}^N)$  the following notation:

$$\langle A \rangle_k = k\text{-vector part of } A, k = 0, \dots, N \quad (2.112)$$

In this notation, the scalar part of  $A$  is given by  $\langle A \rangle_0$ , the vector part  $\langle A \rangle_1$ , the bivector part  $\langle A \rangle_2$ , and so on up to the  $N$ -vector part  $\langle A \rangle_N$ . The parameter  $k$  is said to be the *grade* of  $\langle A \rangle_k$ . Then any multivector in  $\mathcal{G}(\mathbb{R}^N)$  can be written as the sum:

$$A = \langle A \rangle_0 + \langle A \rangle_1 + \dots + \langle A \rangle_N \quad (2.113)$$

If  $A = \langle A \rangle_r$ , i.e.,  $A$  only has an  $r$ -vector part for a given  $r$ , then  $A$  is *homogeneous* and is called an  $r$ -vector. Also, for any integer  $r > 0$ , an  $r$ -vector can be expressed as a sum of  $r$ -blades. The multivector  $A_r$  is an  $r$ -blade (sometimes called a simple  $r$ -vector) if and only if it can be factored into a product of  $r$  *anticommuting* vectors:

$$A_r = \mathbf{a}_1 \mathbf{a}_2 \dots \mathbf{a}_r (= a_1 \wedge a_2 \wedge \dots \wedge a_r) \quad (2.114)$$

where

$$\mathbf{a}_j \mathbf{a}_k = -\mathbf{a}_k \mathbf{a}_j$$

In effect, when  $A_r \neq 0$  the  $r$  vectors  $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_r$  are linearly independent. This follows from the fact that  $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_r$  are mutually orthogonal since they are anticommuting when taken pairwise in the geometric product.

In the material that follows, it will be convenient to introduce the operation of *reversion* of a product of vectors (and hence of a multivector)

$$(\mathbf{a}_1 \mathbf{a}_2 \mathbf{a}_3 \dots \mathbf{a}_r)^\dagger = \mathbf{a}_r \mathbf{a}_{r-1} \mathbf{a}_{r-2} \dots \mathbf{a}_1. \quad (2.115)$$

With respect to multivectors the following relations hold:

$$\begin{aligned} (AB)^\dagger &= B^\dagger A^\dagger \\ (A+B)^\dagger &= A^\dagger + B^\dagger. \\ \langle A^\dagger \rangle &= \langle A \rangle^\dagger \end{aligned} \quad (2.116)$$

As an example, consider the sum of a scalar and a bivector:

$$A = \alpha + \beta \hat{e}_1 \hat{e}_2. \quad (2.117)$$

From the properties of reversion, it follows that

$$\begin{aligned} A^\dagger &= \alpha^\dagger + (\beta \hat{e}_1 \hat{e}_2)^\dagger \\ &= \alpha + \beta \hat{e}_2 \hat{e}_1 \\ &= \alpha - \beta \hat{e}_1 \hat{e}_2. \end{aligned} \quad (2.118)$$

This shows that reversion is analogous to complex conjugation.

When the vectors in a product anticommute (e.g., when they are mutually orthogonal), then by sifting the first vector to the end, sifting the second vector to the position one before the end, and so on, keeping track of sign changes that occur when orthogonal vectors are interchanged, one finds:

$$\langle A^\dagger \rangle_r = \langle A \rangle_r^\dagger = (-1)^{\frac{r(r-1)}{2}} \langle A \rangle_r. \quad (2.119)$$

From these results one may show for an  $r$ -blade:

$$\begin{aligned} A_r A_r^\dagger &= (\mathbf{a}_1 \dots \mathbf{a}_r) (\mathbf{a}_1 \dots \mathbf{a}_r)^\dagger \\ &= |\mathbf{a}_1|^2 \dots |\mathbf{a}_r|^2 \\ &= |A_r|^2, \end{aligned} \quad (2.120)$$

which is a scalar. Note the analogy to contraction rule, which states that the square of a nonzero vector equals a scalar; for vectors  $\mathbf{a}^\dagger = \mathbf{a}$ , so Eq. (2.189) is consistent with this axiom. It follows from Eq. (2.189) that one can define division by a nonzero  $r$ -blade through multiplication by

$$A_r^{-1} = \frac{A_r^\dagger}{|A_r|^2}. \quad (2.121)$$

Thus *it is possible to divide by nonzero  $r$ -blades as well as nonzero vectors*. However, it is not generally true that division by nonzero multivectors is possible. Division by an  $r$ -blade is equivalent to successive divisions by the mutually orthogonal vectors that form the  $r$ -blade.

#### 2.2.2.4 Caution—The Pseudoscalar is Not Simply $\sqrt{-1}$ in Higher Dimensions

As shown before,  $I_2^2 = I_3^2 = -1$ , and it may be tempting from the notation (i.e., denoting the general pseudoscalar with the letter  $I$ , which is similar to  $i$  sometimes used to denote the “imaginary” number  $\sqrt{-1}$ ) that in general  $I_N^2 = -1$ , but the reader is cautioned that this is not true. To understand this point, examine the inverse of the pseudoscalar. It is evident that if the pseudoscalar is represented as

$$I_N = \hat{e}_1 \hat{e}_2 \dots \hat{e}_N \quad (2.122)$$

where  $e_1, e_2, \dots, e_N$  are basis vectors in the  $N$ -dimensional space, then

$$I_N^{-1} = \frac{I_N^\dagger}{|I_N|^2} = \hat{e}_N \hat{e}_{N-1} \dots \hat{e}_1. \quad (2.123)$$

Now, by successively interchanging the unit vectors of  $I_N$  to sift each unit vector backwards through the product and keeping track of the sign changes that occur due to the anticommuting nature of the wedge product, i.e.,

$$\begin{aligned}\hat{\mathbf{e}}_1 \hat{\mathbf{e}}_2 \dots \hat{\mathbf{e}}_{N-1} \hat{\mathbf{e}}_N &= -\hat{\mathbf{e}}_1 \hat{\mathbf{e}}_2 \dots \hat{\mathbf{e}}_N \hat{\mathbf{e}}_{N-1} \\ &= \hat{\mathbf{e}}_1 \hat{\mathbf{e}}_2 \dots \hat{\mathbf{e}}_N \hat{\mathbf{e}}_{N-2} \hat{\mathbf{e}}_{N-1} \\ &= -\hat{\mathbf{e}}_1 \hat{\mathbf{e}}_2 \dots \hat{\mathbf{e}}_N \hat{\mathbf{e}}_{N-3} \hat{\mathbf{e}}_{N-2} \hat{\mathbf{e}}_{N-1}\end{aligned}\quad (2.124)$$

etc., one finds

$$I_N^{-1} = (-1)^{\frac{N(N-1)}{2}} I_N. \quad (2.125)$$

From this it follows that

$$1 = I_N I_N^{-1} = (-1)^{\frac{N(N-1)}{2}} I_N^2. \quad (2.126)$$

Further, because  $N(N-1)$  is always even for  $N \geq 1$  and hence  $(-1)^{N(N-1)} = 1$ , one finally obtains

$$I_N^2 = (-1)^{\frac{N(N-1)}{2}}. \quad (2.127)$$

Thus

$$I_N^2 = \begin{cases} -1 & N(N-1)/2 \text{ odd} \\ 1 & N(N-1)/2 \text{ even} \end{cases}. \quad (2.128)$$

Therefore  $I_N$  is *not* simply a higher dimensional replacement for the unit imaginary  $i = \sqrt{-1}$ .

### 2.2.2.5 Geometric Product of Multivectors

This section examines multiplication of multivectors, thereby extending the earlier work on multiplication of vectors. The results herein are essentially a summary of much more detailed results that can be found in Hestenes and Sobczyk. First consider first the multiplication of a vector and a multivector. As discussed in detail by Hestenes and Sobczyk [8], the geometric product of a vector and a multivector is given by

$$\mathbf{a}A = \mathbf{a} \cdot A + \mathbf{a} \wedge A \quad (2.129)$$

The meanings of the terms on the right-hand side of this expression are elaborated in the following discussion.

First from the notion of grades it follows that the geometric product of vectors yields

$$\begin{aligned}\mathbf{a} \cdot \mathbf{b} &= \langle \mathbf{ab} \rangle_0 \\ \mathbf{a} \wedge \mathbf{b} &= \langle \mathbf{ab} \rangle_2.\end{aligned}\quad (2.130)$$

These relations simply say that the inner product of two vectors, which is a scalar, is given by the grade 0 part of the product  $\mathbf{ab}$ , whereas the outer product of two vectors, which is a bivector, is given by the grade 2 part of the product  $\mathbf{ab}$ . As discussed in Section 2.2.1, it also follows from the definition of the geometric product of vectors that

$$\begin{aligned}\mathbf{a} \cdot \mathbf{b} &= \frac{1}{2}(\mathbf{ab} + \mathbf{ba}) \\ \mathbf{a} \wedge \mathbf{b} &= \frac{1}{2}(\mathbf{ab} - \mathbf{ba})\end{aligned}. \quad (2.131)$$

The notions in Eqs. (2.130) and (2.131) can be used to give meaning to Eq. (2.129).

Consider first the case of  $r$ -blades. By analogy with Eq. (2.130), for an  $r$ -blade  $A_r$ , one may write

$$\begin{aligned}\mathbf{a} \cdot A_r &= \langle \mathbf{a} A_r \rangle_{r-1} \\ \mathbf{a} \wedge A_r &= \langle \mathbf{a} A_r \rangle_{r+1}\end{aligned}. \quad (2.132)$$

When  $A_r$  is a vector, then  $r=1$  and these relations generalize the results in Eq. (2.115). Eq. (2.132) shows that the inner product of a vector with an  $r$ -blade is a *grade-lowering* operation, whereas the wedge product of a vector with an  $r$ -blade is a *grade-raising* operation. Hence, Eq. (2.129) decomposes the geometric product of a vector and an  $r$ -blade into an  $(r-1)$ -grade component and an  $(r+1)$ -grade component.

For the inner product, if  $A_r$  is an  $r$ -blade given by  $A_r = \mathbf{a}_1 \mathbf{a}_2 \dots \mathbf{a}_r$  and  $\mathbf{a} = \alpha \mathbf{a}_i$  where  $\alpha$  is a scalar and  $i \in \{1, 2, \dots, r\}$ , then

$$\mathbf{a} \cdot A_r = \langle \alpha \mathbf{a} \mathbf{a}_1 \mathbf{a}_2 \dots \mathbf{a}_r \rangle_{r-1} = (-1)^{i-1} \alpha |\mathbf{a}_i|^2 \mathbf{a}_2 \wedge \dots \wedge \mathbf{a}_r \quad (2.133)$$

which is an  $r-1$ -vector one grade lower than  $A_r$ . The factor  $(-1)^{i-1}$  arises by sifting  $\mathbf{a}_i$  past  $i-1$  terms to get it next to  $\mathbf{a}_i$  that is already part of  $\mathbf{a}_1 \mathbf{a}_2 \dots \mathbf{a}_r$ . Because these vectors are mutually orthogonal, each such interchange results in a sign change. Once they are beside each other, the contraction rule then yields

$$\mathbf{a}_i \mathbf{a}_i = |\mathbf{a}_i|^2 \quad (2.134)$$

and Eq. (2.133) follows.

If instead one chooses a vector  $\mathbf{a} = \mathbf{a}_{r+1}$  where  $\mathbf{a}_{r+1}$  is mutually orthogonal to each of the vectors comprising the  $r$ -blade, then

$$\mathbf{a} \cdot A_r = 0 \quad (2.135)$$

This occurs because  $\mathbf{a}_{r+1} \mathbf{a}_1 \mathbf{a}_2 \dots \mathbf{a}_r$  is an  $r+1$ -blade (since the vectors entering into this product are mutually orthogonal) and thus writing

$$\mathbf{a}_{r+1} \mathbf{a}_1 \mathbf{a}_2 \dots \mathbf{a}_r = \langle \mathbf{a}_{r+1} \mathbf{a}_1 \mathbf{a}_2 \dots \mathbf{a}_r \rangle_0 + \langle \mathbf{a}_{r+1} \mathbf{a}_1 \mathbf{a}_2 \dots \mathbf{a}_r \rangle_1 + \dots + \langle \mathbf{a}_{r+1} \mathbf{a}_1 \mathbf{a}_2 \dots \mathbf{a}_r \rangle_{r+1} \quad (2.136)$$

yields

$$\langle \mathbf{a}_{r+1} \mathbf{a}_1 \mathbf{a}_2 \dots \mathbf{a}_r \rangle_0 = \langle \mathbf{a}_{r+1} \mathbf{a}_1 \mathbf{a}_2 \dots \mathbf{a}_r \rangle_1 = \dots = \langle \mathbf{a}_{r+1} \mathbf{a}_1 \mathbf{a}_2 \dots \mathbf{a}_r \rangle_r = 0 \quad (2.137)$$

Therefore

$$\mathbf{a} \cdot A_r = \langle \mathbf{a}_{r+1} \mathbf{a}_1 \mathbf{a}_2 \dots \mathbf{a}_r \rangle_{r-1} = 0 \quad (2.138)$$

On the other hand, if one chooses a vector that is a linear combination of some or all of the component vectors, i.e.,

$$\mathbf{a} = \alpha_1 \mathbf{a}_1 + \alpha_2 \mathbf{a}_2 + \dots + \alpha_r \mathbf{a}_r \quad (2.139)$$

then

$$\begin{aligned}\mathbf{a} \cdot A_r &= \langle (\alpha_1 \mathbf{a}_1 + \alpha_2 \mathbf{a}_2 + \cdots + \alpha_r \mathbf{a}_r) \mathbf{a}_1 \mathbf{a}_2 \dots \mathbf{a}_r \rangle_{r-1} \\ &= \alpha_1 \langle \mathbf{a}_1 \mathbf{a}_1 \mathbf{a}_2 \dots \mathbf{a}_r \rangle_{r-1} + \alpha_2 \langle \mathbf{a}_2 \mathbf{a}_1 \mathbf{a}_2 \dots \mathbf{a}_r \rangle_{r-1} + \cdots + \alpha_r \langle \mathbf{a}_r \mathbf{a}_1 \mathbf{a}_2 \dots \mathbf{a}_r \rangle_{r-1}\end{aligned}\quad (2.140)$$

As already discussed before, each of these individual terms is an  $r-1$ -blade and hence the result is  $r-1$ -vector given by a linear combination of  $r-1$ -blades.

Now let  $\mathbf{a} = \alpha \mathbf{a}_1$  and in the product  $\alpha \mathbf{a}_1 \mathbf{a}_2 \dots \mathbf{a}_r$  sift the  $\mathbf{a}$  forward through the products and keep track of the sign changes that occur when orthogonal vectors are interchanged. From the fact that  $\mathbf{a} = \alpha \mathbf{a}_1$  is mutually orthogonal to each of the vectors in the product  $\mathbf{a}_2, \dots, \mathbf{a}_r$  it follows:

$$\begin{aligned}\mathbf{a} \cdot A_r &= \langle \alpha \mathbf{a}_1 \mathbf{a}_1 \mathbf{a}_2 \dots \mathbf{a}_r \rangle_{r-1} \\ &= \langle (-1)^{r-1} \alpha \mathbf{a}_1 \mathbf{a}_2 \dots \mathbf{a}_r, \mathbf{a}_1 \rangle_{r-1} \\ &= \alpha (-1)^{r-1} \langle \mathbf{a}_1 \mathbf{a}_2 \dots \mathbf{a}_r, \mathbf{a}_1 \rangle_{r-1} \\ &= (-1)^{r-1} A_r \cdot \mathbf{a}.\end{aligned}\quad (2.141)$$

This shows that the inner product of a vector with an  $r$ -blade when the vector is parallel to one of the component vectors of the  $r$ -blade is either commutative or anti-commutative, depending on the value of  $r$ . As expected, if  $A_r = A_1$  is a vector, the inner product commutes.

For the outer product, let  $\mathbf{a} = \alpha \mathbf{a}_{r+1}$  where  $\mathbf{a}_{r+1}$  is mutually orthogonal to each of  $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_r$  and examine

$$\mathbf{a} \wedge A_r = \langle \alpha \mathbf{a}_{r+1} \mathbf{a}_1 \mathbf{a}_2 \dots \mathbf{a}_r \rangle_{r+1} = \alpha \mathbf{a}_{r+1} \mathbf{a}_1 \mathbf{a}_2 \dots \mathbf{a}_r. \quad (2.142)$$

This defines a next-higher-grade  $r+1$ -blade. If instead one chooses a vector that is a linear combination of some or all of the component vectors of  $A_r$ , i.e.

$$\mathbf{a} = \alpha_1 \mathbf{a}_1 + \alpha_2 \mathbf{a}_2 + \cdots + \alpha_r \mathbf{a}_r \quad (2.143)$$

then

$$\begin{aligned}\mathbf{a} \wedge A_r &= \langle (\alpha_1 \mathbf{a}_1 + \alpha_2 \mathbf{a}_2 + \cdots + \alpha_r \mathbf{a}_r) \mathbf{a}_1 \mathbf{a}_2 \dots \mathbf{a}_r \rangle_{r+1} \\ &= \alpha_1 \langle \mathbf{a}_1 \mathbf{a}_1 \mathbf{a}_2 \dots \mathbf{a}_r \rangle_{r+1} + \alpha_2 \langle \mathbf{a}_2 \mathbf{a}_1 \mathbf{a}_2 \dots \mathbf{a}_r \rangle_{r+1} + \cdots + \alpha_r \langle \mathbf{a}_r \mathbf{a}_1 \mathbf{a}_2 \dots \mathbf{a}_r \rangle_{r+1} \\ &= 0.\end{aligned}\quad (2.144)$$

This occurs because as already discussed before the general term  $\alpha \mathbf{a}_1 \mathbf{a}_2 \dots \mathbf{a}_r$  leads to an  $r-1$ -blade and hence each term in this sum is equal to 0.

Now sift  $\mathbf{a} = \mathbf{a}_{r+1}$  forward through the products and keep track of the sign changes that occur when orthogonal vectors are interchanged. One finds

$$\begin{aligned}\mathbf{a} \wedge A_r &= \langle (-1)^r \alpha \mathbf{a}_1 \mathbf{a}_2 \dots \mathbf{a}_r, \mathbf{a}_{r+1} \rangle_{r+1} \\ &= \alpha (-1)^r \langle \mathbf{a}_1 \mathbf{a}_2 \dots \mathbf{a}_r, \mathbf{a}_{r+1} \rangle_{r+1} \\ &= (-1)^r A_r \wedge \mathbf{a}.\end{aligned}\quad (2.145)$$

When  $\mathbf{a}$  is mutually orthogonal to each of  $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_r$ , the wedge product is either commutative or anticommutative, depending on the integer  $r$ .

Since an  $r$ -vector can be written as a linear combination of  $r$ -blades, extending the definitions of the inner and outer products of a vector with an  $r$ -blade to definitions of the inner and outer products of a vector with an  $r$ -vector is straightforward.

Now, by analogy with Eq. (2.116) write

$$\begin{aligned}\mathbf{a} \cdot A_r &= \frac{1}{2} (\mathbf{a}A_r + (-1)^{r+1}A_r\mathbf{a}) \\ \mathbf{a} \wedge A_r &= \frac{1}{2} (\mathbf{a}A_r - (-1)^{r+1}A_r\mathbf{a})\end{aligned}. \quad (2.146)$$

These relations give the proper commutation behavior as discussed before. For example, when  $r=1$  and hence  $A_1$  is a vector, Eq. (2.146) reduces to Eq. (2.131).

It is useful to note that in addition to reproducing the previous results on the multiplication of vectors, setting  $r=0$  in Eq. (2.146) shows that the inner product of a vector and scalar equals 0, whereas the outer product of a vector and a scalar equals the scalar times the vector, which commutes. Thus by Eq. (2.139), the product of a scalar and a vector commutes, as expected.

By analogy with Eq. (2.132), one can now extend these products to general  $r$ -vectors by

$$\begin{aligned}A_r \cdot B_s &= \langle A_r B_s \rangle_{|r-s|} \\ A_r \wedge B_s &= \langle A_r B_s \rangle_{r+s}\end{aligned}. \quad (2.147)$$

and to general multivectors comprising sums of  $r$ -vectors, by summation

$$\begin{aligned}A \cdot B &= \sum A_r \cdot B_s \\ A \wedge B &= \sum A_r \wedge B_s\end{aligned}. \quad (2.148)$$

It is important to note that, unlike in the development before, for general multivectors

$$AB \neq A \cdot B + A \wedge B \quad (2.149)$$

There may be additional terms that arise in the product of two general multivectors. Thus the relation in Eq. (2.129) holds only for the product of a vector and a multivector. The full generality of geometric algebra and the products of multivectors will not be used herein, but the interested reader is encouraged to consult Hestenes and Sobczyk for a much more complete discussion of these ideas.

It is now of some interest to examine the quantity  $A_r \mathbf{a}$ . The relations in Eq. (2.146) imply

$$\begin{aligned}A_r \mathbf{a} &= (-1)^{r+1} 2 \mathbf{a} \cdot A_r - (-1)^{r+1} \mathbf{a} A_r \\ A_r \mathbf{a} &= (-1)^{r+1} \mathbf{a} A_r - 2(-1)^{r+1} \mathbf{a} \wedge A_r\end{aligned}. \quad (2.150)$$

Combining these two relations implies

$$\begin{aligned}A_r \mathbf{a} &= (-1)^{r+1} \mathbf{a} \cdot A_r - \frac{1}{2} (-1)^{r+1} \mathbf{a} A_r + \frac{1}{2} (-1)^{r+1} \mathbf{a} A_r - (-1)^{r+1} \mathbf{a} \wedge A_r \\ &= (-1)^{r+1} (\mathbf{a} \cdot A_r - \mathbf{a} \wedge A_r)\end{aligned}. \quad (2.151)$$

As an example, if  $A_r$  is a bivector with  $r=2$ , this result becomes

$$A_2 \mathbf{a} = -\mathbf{a} \cdot A_r + \mathbf{a} \wedge A_r \quad (2.152)$$

The distinction between  $\mathbf{a}A_r$  and  $A_r\mathbf{a}$  is important and must be kept in mind.

An important example of the product of a vector with a multivector is the inner product of a vector with a bivector. Assume a vector space  $\mathbb{R}^N$  with orthonormal basis vectors  $\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \dots, \hat{\mathbf{e}}_N$  and let  $\mathbf{a}, \mathbf{b}, \mathbf{c}$  be vectors in  $\mathbb{R}^N$ . Examine the relation

$$\begin{aligned}\mathbf{a} \cdot (\mathbf{b} \wedge \mathbf{c}) &= \frac{1}{4} \mathbf{a}(\mathbf{bc} - \mathbf{cb}) - \frac{1}{4}(\mathbf{bc} - \mathbf{cb})\mathbf{a} \\ &= \frac{1}{4}(\mathbf{abc} + \mathbf{cba}) - \frac{1}{4}(\mathbf{acb} + \mathbf{bca})\end{aligned}\quad (2.153)$$

Adding and subtracting terms  $\mathbf{bac}$  and  $\mathbf{cab}$  does not change this relation:

$$\begin{aligned}\mathbf{a} \cdot (\mathbf{b} \wedge \mathbf{c}) &= \frac{1}{4}(\mathbf{abc} + \mathbf{bac} + \mathbf{cba} - \mathbf{bac}) - \frac{1}{4}(\mathbf{acb} + \mathbf{cab} + \mathbf{bca} - \mathbf{cab}) \\ &= \frac{1}{4}((\mathbf{ab} + \mathbf{ba})\mathbf{c} + \mathbf{c}(\mathbf{ba} + \mathbf{ab}) - (\mathbf{ac} + \mathbf{ca})\mathbf{b} - \mathbf{b}(\mathbf{ac} + \mathbf{ca}))\end{aligned}\quad (2.154)$$

By Eq. (2.131) this yields

$$\mathbf{a} \cdot (\mathbf{b} \wedge \mathbf{c}) = (\mathbf{a} \cdot \mathbf{b})\mathbf{c} - (\mathbf{a} \cdot \mathbf{c})\mathbf{b} \quad (2.155)$$

This is a vector in the two-dimensional subspace determined by the vectors  $\mathbf{b}$  and  $\mathbf{c}$ . Thus the vector  $\mathbf{a} \cdot (\mathbf{b} \wedge \mathbf{c})$  may be thought of as a kind of projection of  $\mathbf{a}$  into the subspace determined by the vectors  $\mathbf{b}$  and  $\mathbf{c}$ . However, it is not an *orthogonal* projection.

To further examine this notion of projection, let unit vectors  $\hat{\mathbf{e}}_1$  and  $\hat{\mathbf{e}}_2$  be orthogonal. Thus they form an orthonormal basis for a two-dimensional subspace of  $\mathbb{R}^N$  (where obviously  $N \geq 2$ ). Examine the inner product of a vector  $\mathbf{a}$  with the unit bivector  $\hat{\mathbf{e}}_1 \wedge \hat{\mathbf{e}}_2$ :

$$\mathbf{a} \cdot (\hat{\mathbf{e}}_1 \wedge \hat{\mathbf{e}}_2) = (\mathbf{a} \cdot \hat{\mathbf{e}}_1)\hat{\mathbf{e}}_2 - (\mathbf{a} \cdot \hat{\mathbf{e}}_2)\hat{\mathbf{e}}_1 \quad (2.156)$$

As already observed, this is a vector in the subspace determined by the unit vectors  $\hat{\mathbf{e}}_1$  and  $\hat{\mathbf{e}}_2$ . Note also that because  $\hat{\mathbf{e}}_1$  and  $\hat{\mathbf{e}}_2$  are orthogonal, it is also true that

$$\hat{\mathbf{e}}_1 \wedge \hat{\mathbf{e}}_2 = \hat{\mathbf{e}}_1 \hat{\mathbf{e}}_2 \quad (2.157)$$

It then follows that the inverse of the unit bivector  $\hat{\mathbf{e}}_1 \wedge \hat{\mathbf{e}}_2$  is given by

$$(\hat{\mathbf{e}}_1 \wedge \hat{\mathbf{e}}_2)^{-1} = \hat{\mathbf{e}}_2 \hat{\mathbf{e}}_1 \quad (2.158)$$

Using the contraction rule and noting that  $\hat{\mathbf{e}}_1$  and  $\hat{\mathbf{e}}_2$  are unit vectors, the reader can easily confirm that

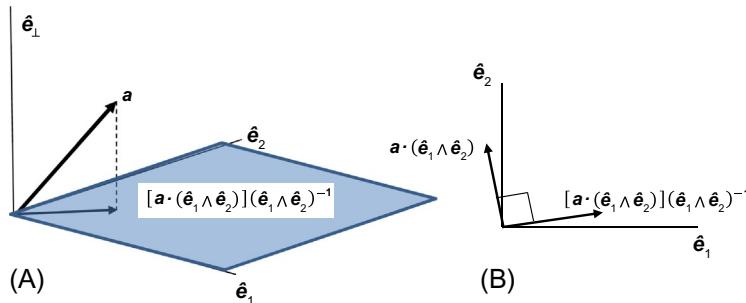
$$(\hat{\mathbf{e}}_1 \wedge \hat{\mathbf{e}}_2)^{-1}(\hat{\mathbf{e}}_1 \wedge \hat{\mathbf{e}}_2) = (\hat{\mathbf{e}}_1 \wedge \hat{\mathbf{e}}_2)(\hat{\mathbf{e}}_1 \wedge \hat{\mathbf{e}}_2)^{-1} = 1 \quad (2.159)$$

Now examine the quantity

$$\begin{aligned}[\mathbf{a} \cdot (\hat{\mathbf{e}}_1 \wedge \hat{\mathbf{e}}_2)](\hat{\mathbf{e}}_1 \wedge \hat{\mathbf{e}}_2)^{-1} &= [(\mathbf{a} \cdot \hat{\mathbf{e}}_1)\hat{\mathbf{e}}_2 + (\mathbf{a} \cdot \hat{\mathbf{e}}_2)\hat{\mathbf{e}}_1]\hat{\mathbf{e}}_2 \hat{\mathbf{e}}_1 \\ &= (\mathbf{a} \cdot \hat{\mathbf{e}}_1)\hat{\mathbf{e}}_1 + (\mathbf{a} \cdot \hat{\mathbf{e}}_2)\hat{\mathbf{e}}_2\end{aligned}\quad (2.160)$$

From Eq. (2.160) it is apparent that  $[\mathbf{a} \cdot (\hat{\mathbf{e}}_1 \wedge \hat{\mathbf{e}}_2)](\hat{\mathbf{e}}_1 \wedge \hat{\mathbf{e}}_2)^{-1}$  is the *orthogonal projection* of the vector  $\mathbf{a}$  into the subspace defined by the unit vectors  $\hat{\mathbf{e}}_1$  and  $\hat{\mathbf{e}}_2$ . Moreover,

$$[(\mathbf{a} \cdot \hat{\mathbf{e}}_1)\hat{\mathbf{e}}_2 - (\mathbf{a} \cdot \hat{\mathbf{e}}_2)\hat{\mathbf{e}}_1] \cdot [(\mathbf{a} \cdot \hat{\mathbf{e}}_1)\hat{\mathbf{e}}_1 + (\mathbf{a} \cdot \hat{\mathbf{e}}_2)\hat{\mathbf{e}}_2] = -(\mathbf{a} \cdot \hat{\mathbf{e}}_2)(\mathbf{a} \cdot \hat{\mathbf{e}}_1) + (\mathbf{a} \cdot \hat{\mathbf{e}}_1)(\mathbf{a} \cdot \hat{\mathbf{e}}_2) = 0 \quad (2.161)$$

**FIG. 2.20**

(A) Orthogonal projection of a vector onto a bivector in geometric algebra; (B) relation between two projected vectors.

Thus these two vectors are orthogonal to each other. The relations between  $a \cdot (\hat{e}_1 \wedge \hat{e}_2)$  and  $[a \cdot (\hat{e}_1 \wedge \hat{e}_2)](\hat{e}_1 \wedge \hat{e}_2)^{-1}$  are illustrated in Fig. 2.20, in which  $\hat{e}_\perp$  signifies the remainder of the space orthogonal to  $\hat{e}_1$  and  $\hat{e}_2$ .

This is an important result in the context of signal processing as engineers project vectors into subspaces quite often. An approach to projecting vectors into more general subspaces will be discussed further in Section 2.2.4.

### 2.2.3 WHAT IS A COMPLEX NUMBER?

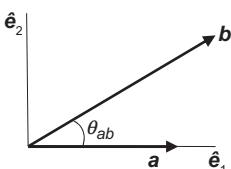
To explore the geometric properties of the product of two vectors in a Euclidean plane, choose a particular orthonormal basis as follows. Let

$$\hat{e}_1 = \frac{\mathbf{a}}{|\mathbf{a}|} \quad (2.162)$$

and let  $\hat{e}_2$  be orthogonal (in the plane containing  $\mathbf{a}$  and  $\mathbf{b}$ ) to  $\hat{e}_1$ . Let  $\mathbf{b}$  be given in this plane by

$$\begin{aligned} \mathbf{b} &= (\mathbf{b} \cdot \hat{e}_1)\hat{e}_1 + (\mathbf{b} \cdot \hat{e}_2)\hat{e}_2 \\ &= \frac{(\mathbf{b} \cdot \mathbf{a})}{|\mathbf{a}|^2}\mathbf{a} + (\mathbf{b} \cdot \hat{e}_2)\hat{e}_2 \end{aligned} \quad (2.163)$$

These vectors are illustrated in Fig. 2.21:

**FIG. 2.21**

Coplanar vectors.

It is evident from this picture that

$$b_1 = (\mathbf{b} \cdot \hat{\mathbf{e}}_1) = |\mathbf{b}| \cos \theta_{ab} \quad (2.164)$$

$$b_2 = (\mathbf{b} \cdot \hat{\mathbf{e}}_2) = |\mathbf{b}| \sin \theta_{ab} \quad (2.165)$$

$$a_1 = (\mathbf{a} \cdot \hat{\mathbf{e}}_1) = |\mathbf{a}| \quad (2.166)$$

$$a_2 = (\mathbf{a} \cdot \hat{\mathbf{e}}_2) = 0 \quad (2.167)$$

where  $\theta_{ab}$  is the angle between  $\mathbf{a}$  and  $\mathbf{b}$  in their shared plane. Hence

$$\begin{aligned} \mathbf{a}\mathbf{b} &= \mathbf{a} \cdot \mathbf{b} + \mathbf{a} \wedge \mathbf{b} \\ &= a_1 b_1 + a_2 b_2 + (a_1 b_2 - a_2 b_1) \hat{\mathbf{e}}_1 \hat{\mathbf{e}}_2 \\ &= |\mathbf{a}| |\mathbf{b}| \cos \theta_{ab} \pm |\mathbf{a}| |\mathbf{b}| \sin \theta_{ab} \hat{\mathbf{e}}_1 \hat{\mathbf{e}}_2 \end{aligned} \quad (2.168)$$

The sign is determined by the sign of  $(a_1 b_2 - a_2 b_1)$ . The individual inner and outer products are given by

$$\mathbf{a} \cdot \mathbf{b} = |\mathbf{a}| |\mathbf{b}| \cos \theta_{ab} \quad (2.169)$$

$$\mathbf{a} \wedge \mathbf{b} = \pm |\mathbf{a} \wedge \mathbf{b}| \hat{\mathbf{e}}_1 \hat{\mathbf{e}}_2 = \pm |\mathbf{a}| |\mathbf{b}| \sin \theta_{ab} \hat{\mathbf{e}}_1 \hat{\mathbf{e}}_2 \quad (2.170)$$

Note here that  $\sin \theta_{ab} \geq 0$  (hence  $0 \leq \theta_{ab} \leq \pi$ ). In terms of the magnitudes of the vectors and the angles between them the geometric product of  $\mathbf{a}$  and  $\mathbf{b}$  becomes

$$\mathbf{a}\mathbf{b} = |\mathbf{a}| |\mathbf{b}| \cos \theta_{ab} + I_{ab} |\mathbf{a}| |\mathbf{b}| \sin \theta_{ab} \quad (2.171)$$

Because algebraically  $I_{ab}^2 = -1$ , one may define the following function of a scalar  $x$ :

$$e^{I_{ab}x} \triangleq \sum_{k=0}^{\infty} \frac{(I_{ab}x)^k}{k!} \quad (2.172)$$

Using the associativity of the product of vectors this summation may be expanded as

$$\begin{aligned} e^{I_{ab}x} &= \sum_{m=0}^{\infty} \frac{(I_{ab})^m x^{2m}}{2m!} + I_{ab} \sum_{m=0}^{\infty} \frac{(I_{ab})^m x^{2m+1}}{(2m+1)!} \\ &= \cos x + I_{ab} \sin x \end{aligned} \quad (2.173)$$

This result is a version of Euler's formula that incorporates the unit bivector  $I_{ab}$ . With this result it follows

$$\mathbf{a}\mathbf{b} = |\mathbf{a}| |\mathbf{b}| e^{I_{ab}\theta_{ab}}. \quad (2.174)$$

Similarly

$$\mathbf{b}\mathbf{a} = (\mathbf{a}\mathbf{b})^\dagger = |\mathbf{a}| |\mathbf{b}| e^{-I_{ab}\theta_{ab}} \quad (2.175)$$

The relationship between a two-dimensional vector and a corresponding complex number follows from

$$\begin{aligned} \hat{\mathbf{e}}_1 \mathbf{a} &= \hat{\mathbf{e}}_1 a_1 \hat{\mathbf{e}}_1 + \hat{\mathbf{e}}_1 a_2 \hat{\mathbf{e}}_2 \\ &= a_1 \hat{\mathbf{e}}_1 \hat{\mathbf{e}}_1 + a_2 \hat{\mathbf{e}}_1 \hat{\mathbf{e}}_2 \\ &= a_1 + I_{\hat{\mathbf{e}}_1 \hat{\mathbf{e}}_2} a_2 \end{aligned} \quad (2.176)$$

$$\begin{aligned} (\hat{\mathbf{e}}_1 \mathbf{a})^\dagger &= \mathbf{a} \hat{\mathbf{e}}_1 = \hat{\mathbf{e}}_1 a_1 \hat{\mathbf{e}}_1 + a_2 \hat{\mathbf{e}}_2 \hat{\mathbf{e}}_1 \\ &= a_1 - I_{\hat{\mathbf{e}}_1 \hat{\mathbf{e}}_2} a_2 \end{aligned} \quad (2.177)$$

Thus if  $\hat{\mathbf{e}}_1 \mathbf{a}$  is the “complex number,” then  $(\hat{\mathbf{e}}_1 \mathbf{a})^\dagger$  is its complex conjugate. In this way one can easily switch between two-dimensional vectors and complex numbers, but nonetheless maintain a distinct view of these two very different kinds of quantities. Note also that the particular complex number associated with a vector is a function of  $\hat{\mathbf{e}}_1$ , i.e., it depends on which axis is designated to play the role of the “real” axis.

### 2.2.3.1 Rotation of Vectors via Spinors

Using the geometric product defined before now examine the product of three vectors that all share the same plane:

$$\mathbf{abc}$$

To perform this computation, set  $\mathbf{c} = c_1 \hat{\mathbf{e}}_1 + c_2 \hat{\mathbf{e}}_2$  and use the geometric product for  $\mathbf{ab}$  to write

$$\begin{aligned} \mathbf{abc} &= |\mathbf{a}| |\mathbf{b}| e^{I_{ab} \theta_{ab}} \mathbf{c} \\ &= |\mathbf{a}| |\mathbf{b}| (\cos \theta_{ab} + I_{ab} \sin \theta_{ab}) \mathbf{c} \end{aligned} \quad (2.178)$$

Here we have also used the Euler representation derived before. To complete this computation, examine

$$\begin{aligned} I_{ab} \mathbf{c} &= \hat{\mathbf{e}}_1 \hat{\mathbf{e}}_2 (c_1 \hat{\mathbf{e}}_1 + c_2 \hat{\mathbf{e}}_2) \\ &= c_1 \hat{\mathbf{e}}_1 \hat{\mathbf{e}}_2 \hat{\mathbf{e}}_1 + c_2 \hat{\mathbf{e}}_1 \hat{\mathbf{e}}_2 \hat{\mathbf{e}}_2 \\ &= -c_1 \hat{\mathbf{e}}_2 + c_2 \hat{\mathbf{e}}_1 \end{aligned} \quad (2.179)$$

Hence

$$\begin{aligned} \mathbf{abc} &= |\mathbf{a}| |\mathbf{b}| (\cos \theta_{ab} \mathbf{c} + (-c_1 \hat{\mathbf{e}}_2 + c_2 \hat{\mathbf{e}}_1) \sin \theta_{ab}) \\ &= |\mathbf{a}| |\mathbf{b}| \{(c_1 \cos \theta_{ab} + c_2 \sin \theta_{ab}) \hat{\mathbf{e}}_1 + (-c_1 \sin \theta_{ab} + c_2 \cos \theta_{ab}) \hat{\mathbf{e}}_2\} \end{aligned} \quad (2.180)$$

This is a vector in the same  $\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2$  plane. In fact it is simply a rotation of  $\mathbf{c}$  by the angle  $\theta_{ab}$  and a dilation by  $|\mathbf{a}| |\mathbf{b}|$ . This is the result we expect if we view the product of two vectors  $\mathbf{ab}$  as a “complex” number. Thus the product of three coplanar vectors is another vector in the same plane. Consequently, the product  $\mathbf{ab}$  may be viewed as an *operator* that converts a vector  $\mathbf{c}$  in the  $\mathbf{a}, \mathbf{b}$  plane into another vector in the same plane. The product  $\mathbf{ab}$  is an example of a *spinor*.

The spinor formulation provides a very convenient framework for studying rotations of vectors. Again consider the product

$$\mathbf{abc}$$

However, now let the vector  $\mathbf{c}$  have both coplanar and nonco-planar components:

$$\mathbf{c} = \mathbf{c}_{\parallel} + \mathbf{c}_{\perp} \quad (2.181)$$

In this expression  $\mathbf{c}_{\parallel}$  is in the  $\mathbf{a}, \mathbf{b}$  plane and  $\mathbf{c}_{\perp}$  is orthogonal to this plane. Thus

$$\mathbf{abc} = \mathbf{abc}_{\parallel} + \mathbf{abc}_{\perp} \quad (2.182)$$

The previous considerations show that  $\mathbf{abc}_{\parallel}$  is a vector in the  $\mathbf{a}, \mathbf{b}$  plane. In particular,  $\mathbf{abc}_{\parallel}$  is a dilation and rotation of the part of  $\mathbf{c}$  that is *in the plane*.

Now examine  $\mathbf{abc}_\perp$ :

$$\begin{aligned}\mathbf{abc}_\perp &= |\mathbf{a}||\mathbf{b}|(\cos\theta_{ab} + I_{ab}\sin\theta_{ab})\mathbf{c}_\perp \\ &= |\mathbf{a}||\mathbf{b}|(\cos\theta_{ab}\mathbf{c}_\perp + I_{ab}\mathbf{c}_\perp \sin\theta_{ab})\end{aligned}\quad (2.183)$$

To continue it is necessary to determine the value of  $I_{ab}\mathbf{c}_\perp$ . Write

$$I_{ab}\mathbf{c}_\perp = \hat{\mathbf{e}}_1 \hat{\mathbf{e}}_2 \mathbf{c}_\perp \quad (2.184)$$

Observe now that  $\mathbf{c}_\perp$  is by assumption mutually orthogonal with  $\hat{\mathbf{e}}_1$  and  $\hat{\mathbf{e}}_2$ . Thus  $\hat{\mathbf{e}}_1 \hat{\mathbf{e}}_2 \mathbf{c}_\perp$  defines a trivector. As a result  $\mathbf{abc}_\perp$  is the sum of a vector and a trivector. Combining these two results yields

$$\mathbf{abc} = \mathbf{abc}_\parallel + (|\mathbf{a}||\mathbf{b}|\cos\theta_{ab}\mathbf{c}_\perp + |\mathbf{a}||\mathbf{b}|\sin\theta_{ab}I_{ab}\mathbf{c}_\perp) \quad (2.185)$$

The product  $\mathbf{abc}$  where  $\mathbf{c}$  has components that are orthogonal to the  $\mathbf{a}, \mathbf{b}$  plane results in the *sum of a vector and a trivector*. This clearly then does not lead to a rotation of  $\mathbf{c}$  in the  $\mathbf{a}, \mathbf{b}$  plane.

To continue, write

$$\mathbf{c}_\perp = c_\perp \hat{\mathbf{e}}_\perp \quad (2.186)$$

where  $\hat{\mathbf{e}}_\perp$  is mutually orthogonal with  $\hat{\mathbf{e}}_1$  and  $\hat{\mathbf{e}}_2$ . Because the three unit vectors are mutually orthogonal, one may sift  $\hat{\mathbf{e}}_\perp$  through the product, changing signs as necessary, to write

$$I_{ab}\hat{\mathbf{e}}_\perp = \hat{\mathbf{e}}_1 \hat{\mathbf{e}}_2 \hat{\mathbf{e}}_\perp = -\hat{\mathbf{e}}_1 \hat{\mathbf{e}}_\perp \hat{\mathbf{e}}_2 = \hat{\mathbf{e}}_\perp \hat{\mathbf{e}}_1 \hat{\mathbf{e}}_2 = \hat{\mathbf{e}}_\perp I_{ab} \quad (2.187)$$

Since  $\hat{\mathbf{e}}_\perp$  is orthogonal to the  $\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2$  plane, which is the same as the  $\mathbf{a}, \mathbf{b}$  plane, but is otherwise arbitrary, it follows that *the unit bivector  $I_{ab}$  commutes with all vectors that are orthogonal to the  $\mathbf{a}, \mathbf{b}$  plane*. This property of  $I_{ab}$  permits a very convenient characterization of rotations through the operation of a spinor. In particular, examine a *two-sided* spinor operation

$$\mathbf{abcba} = |\mathbf{a}|^2 |\mathbf{b}|^2 e^{I_{ab}\theta_{ab}} \mathbf{c} e^{-I_{ab}\theta_{ab}} \quad (2.188)$$

This expression uses the Euler representations of the products  $\mathbf{ab}$  and  $\mathbf{ba}$  that were obtained before.

First consider  $\mathbf{c}_\parallel$  in the plane:

$$e^{I_{ab}\theta_{ab}} \mathbf{c}_\parallel e^{-I_{ab}\theta_{ab}} \quad (2.189)$$

The previous considerations have already shown

$$e^{I_{ab}\theta_{ab}} \mathbf{c}_\parallel = c_{rotated,1} \hat{\mathbf{e}}_1 + c_{rotated,2} \hat{\mathbf{e}}_2 \quad (2.190)$$

where

$$c_{rotated,1} = c_1 \cos\theta_{ab} + c_2 \sin\theta_{ab} \quad (2.191)$$

$$c_{rotated,2} = -c_1 \sin\theta_{ab} + c_2 \cos\theta_{ab} \quad (2.192)$$

Continuing computation of the two-sided operation now yields

$$\begin{aligned}e^{I_{ab}\theta_{ab}} \mathbf{c}_\parallel e^{-I_{ab}\theta_{ab}} &= \{c_{rotated,1} \hat{\mathbf{e}}_1 + c_{rotated,2} \hat{\mathbf{e}}_2\} e^{-I_{ab}\theta_{ab}} \\ &= \{c_{rotated,1} \hat{\mathbf{e}}_1 + c_{rotated,2} \hat{\mathbf{e}}_2\} \{\cos\theta_{ab} - \sin\theta_{ab} I_{ab}\} \\ &= [c_{rotated,1} \cos\theta_{ab} + c_{rotated,2} \sin\theta_{ab}] \hat{\mathbf{e}}_1 + [-c_{rotated,1} \sin\theta_{ab} + c_{rotated,2} \cos\theta_{ab}] \hat{\mathbf{e}}_2\end{aligned}\quad (2.193)$$

This again is simply a planar rotation of the vector  $c_{rotated,1}\hat{\mathbf{e}}_1 + c_{rotated,2}\hat{\mathbf{e}}_2$ , which already incorporates a rotation of the planar part of  $\mathbf{c}$  by an angle  $\theta_{ab}$ , by an *additional* angle  $\theta_{ab}$ . Hence the two-sided operation rotates the coplanar vector  $\mathbf{c}_{\parallel}$  in the plane by the angle  $2\theta_{ab}$ .

At this point it is legitimate to ask why one needs a two-sided operation to obtain a rotation of  $2\theta_{ab}$  instead of a one-sided operation to obtain a rotation of  $\theta_{ab}$ . Recall from before that the one-sided operation of  $\mathbf{ab}$  on  $\mathbf{c}$  yields the sum of a vector and a trivector, with the trivector arising due to the part  $\mathbf{c}_{\perp}$  of  $\mathbf{c}$  that is not in the  $\mathbf{a}, \mathbf{b}$ -plane. *The two-side operation allows us to avoid this trivector.* Examine:

$$\begin{aligned} e^{I_{ab}\theta_{ab}}\mathbf{c}_{\perp}e^{-I_{ab}\theta_{ab}} &= \{\cos\theta_{ab} + \sin\theta_{ab}I_{ab}\}\mathbf{c}_{\perp}\{\cos\theta_{ab} - \sin\theta_{ab}I_{ab}\} \\ &= \{\cos^2\theta_{ab}\mathbf{c}_{\perp} + \sin\theta_{ab}\cos\theta_{ab}I_{ab}\mathbf{c}_{\perp} - \cos\theta_{ab}\sin\theta_{ab}\mathbf{c}_{\perp}I_{ab} - \sin^2\theta_{ab}I_{ab}\mathbf{c}_{\perp}I_{ab}\} \end{aligned} \quad (2.194)$$

As we discussed before,  $I_{ab}$  commutes with any vector that is orthogonal to the  $\mathbf{a}, \mathbf{b}$ -plane—hence it commutes with  $\mathbf{c}_{\perp}$ . Therefore the second and third terms in this expression yield

$$\sin\theta_{ab}\cos\theta_{ab}I_{ab}\mathbf{c}_{\perp} - \cos\theta_{ab}\sin\theta_{ab}\mathbf{c}_{\perp}I_{ab} = 0 \quad (2.195)$$

In addition, the final term yields

$$-\sin^2\theta_{ab}I_{ab}\mathbf{c}_{\perp}I_{ab} = -\sin^2\theta_{ab}I_{ab}^2\mathbf{c}_{\perp} = \sin^2\theta_{ab}\mathbf{c}_{\perp} \quad (2.196)$$

Therefore the result becomes

$$\begin{aligned} e^{I_{ab}\theta_{ab}}\mathbf{c}_{\perp}e^{-I_{ab}\theta_{ab}} &= \{\cos^2\theta_{ab} + \sin^2\theta_{ab}\}\mathbf{c}_{\perp} \\ &= \mathbf{c}_{\perp} \end{aligned} \quad (2.197)$$

*The two-sided operation does not rotate vectors that are orthogonal to the  $\mathbf{a}, \mathbf{b}$ -plane.* As a result it avoids the trivector that arises from the one-side operation. Therefore we find in general

$$\mathbf{abeba} = |\mathbf{a}|^2|\mathbf{b}|^2(\mathbf{c}_{\parallel,rotated} + \mathbf{c}_{\perp}) \quad (2.198)$$

with

$$\mathbf{c}_{\parallel,rotated} = e^{I_{ab}\theta_{ab}}\mathbf{c}_{\parallel}e^{-I_{ab}\theta_{ab}} \quad (2.199)$$

The previous considerations provide a very convenient approach to rotations. In particular define the following *rotors*, which are unit spinors obtained as the product of two unit vectors:

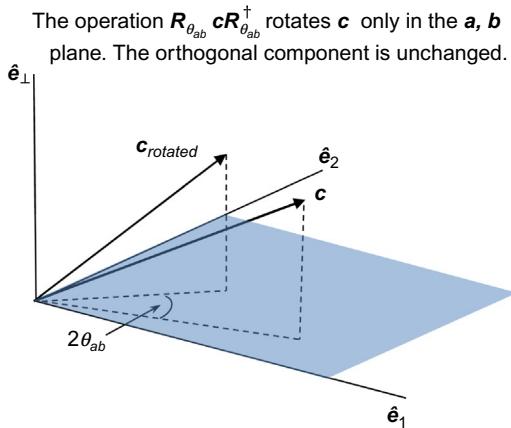
$$R_{\theta_{ab}} \triangleq \hat{\mathbf{a}}\hat{\mathbf{b}} = e^{I_{ab}\theta_{ab}} \quad (2.200)$$

$$R_{\theta_{ab}}^\dagger \triangleq \hat{\mathbf{b}}\hat{\mathbf{a}} = e^{-I_{ab}\theta_{ab}} \quad (2.201)$$

In these expressions  $I_{ab}$  is the unit bivector in the  $\mathbf{a}, \mathbf{b}$ -plane with the same orientation as  $\mathbf{a} \wedge \mathbf{b}$ . The two-sided spinor operation is now expressed as

$$R_{\theta_{ab}}\mathbf{c}R_{\theta_{ab}}^\dagger = \mathbf{c}_{\parallel,rotated} + \mathbf{c}_{\perp} \quad (2.202)$$

This operation rotates the part of the vector  $\mathbf{c}$  in the  $\mathbf{a}, \mathbf{b}$  plane through an angle  $2\theta_{ab}$ , but leaves the part of the vector  $\mathbf{c}$  orthogonal to the plane untouched. As long as one operates only on coplanar vectors, rotation can be defined via a one-sided operation.

**FIG. 2.22**


---

Rotation using rotors.

However, the one-sided operation no longer yields only a rotated vector when applied to vectors with components orthogonal to the plane of interest. In this more general case, rotation by an angle  $\theta$  in a given plane can be implemented conveniently via the two-side rotor operation with rotors that each implement half of the desired rotation. This notion is illustrated in Fig. 2.22.

## 2.2.4 WHAT IS A COMPLEX VECTOR?

### 2.2.4.1 N-Dimensional Complex Vector as a 2N-Dimensional Real Vector

Consider a sinusoidal signal

$$s(t) = A \cos((\omega_0 + \omega_\Delta)t + \theta). \quad (2.203)$$

Assume a known nominal frequency  $\omega_0$  and let  $\omega_\Delta$  be an offset frequency. Simple sum and difference identities from trigonometry lead to

$$s(t) = \mathcal{J}_{\Delta,\theta}(t) \cos \omega_0 t - \mathcal{Q}_{\Delta,\theta}(t) \sin \omega_0 t \quad (2.204)$$

where

$$\mathcal{J}_{\Delta,\theta}(t) = A[\cos \theta \cos \omega_\Delta t - \sin \theta \sin \omega_\Delta t] \quad (2.205)$$

$$\mathcal{Q}_{\Delta,\theta}(t) = A[\cos \theta \sin \omega_\Delta t + \sin \theta \cos \omega_\Delta t] \quad (2.206)$$

Because  $\omega_0$  is known, one can recover  $s(t)$  from knowledge of  $\mathcal{J}_{\Delta,\theta}(t)$  and  $\mathcal{Q}_{\Delta,\theta}(t)$ . In this sense the pair  $(\mathcal{J}_{\Delta,\theta}(t), \mathcal{Q}_{\Delta,\theta}(t))$  represents the signal  $s(t)$ .

Sampling the signal at times  $t = t_i$ ,  $i = 1, \dots, N$  yields two sets of samples

$$\begin{aligned} \mathcal{J}_{\Delta,\theta,i} &= \mathcal{J}_{\Delta,\theta}(t_i) & i = 1, \dots, N. \\ \mathcal{Q}_{\Delta,\theta,i} &= \mathcal{Q}_{\Delta,\theta}(t_i) \end{aligned} \quad (2.207)$$

One may assume  $t_1 = 0$  without loss of generality. A signal sample  $s(t_i)$  is described by

$$s(t_i) = d_{\Delta,\theta,i} b_{0,i} + d_{\Delta,\theta,i+N} b_{0,i+N}. \quad (2.208)$$

where

$$\mathbf{d}_{\Delta,\theta} = \begin{bmatrix} \mathcal{J}_{\Delta,\theta,1} \\ \vdots \\ \mathcal{J}_{\Delta,\theta,N} \\ \mathcal{Q}_{\Delta,\theta,1} \\ \vdots \\ \mathcal{Q}_{\Delta,\theta,N} \end{bmatrix} \quad (2.209)$$

$$\mathbf{b}_\theta = \begin{bmatrix} \cos \omega_0 t_1 \\ \vdots \\ \cos \omega_0 t_N \\ -\sin \omega_0 t_1 \\ \vdots \\ -\sin \omega_0 t_N \end{bmatrix} \quad (2.210)$$

The  $2N$ -dimensional real vector  $\mathbf{d}_{\Delta,\theta}$  represents the entire set of  $N$  sinusoidal signal samples  $s(t_i), i=1, \dots, N$ . It is important to observe that each frequency offset  $\omega_\Delta$  and each phase offset  $\theta$  gives rise to a different vector  $\mathbf{d}_{\Delta,\theta}$ .

For a given frequency  $\omega_\Delta$  define the following vectors:

$$\hat{\mathbf{p}}_\Delta = \frac{1}{\sqrt{N}} \begin{bmatrix} \cos \omega_\Delta t_1 \\ \vdots \\ \cos \omega_\Delta t_N \\ \sin \omega_\Delta t_1 \\ \vdots \\ \sin \omega_\Delta t_N \end{bmatrix}; \quad \hat{\mathbf{p}}'_\Delta = \frac{1}{\sqrt{N}} \begin{bmatrix} -\sin \omega_\Delta t_1 \\ \vdots \\ -\sin \omega_\Delta t_N \\ \cos \omega_\Delta t_1 \\ \vdots \\ \cos \omega_\Delta t_N \end{bmatrix} \quad (2.211)$$

These vectors have unit magnitude ( $|\hat{\mathbf{p}}_\Delta|^2 = |\hat{\mathbf{p}}'_\Delta|^2 = 1$ ) and are orthogonal ( $\hat{\mathbf{p}}_\Delta \cdot \hat{\mathbf{p}}'_\Delta = 0$ ) as is easily shown. Therefore they form an orthonormal basis for  $\mathbf{d}_{\Delta,\theta}$ :

$$\mathbf{d}_{\Delta,\theta} = A \cos \theta \hat{\mathbf{p}}_\Delta + A \sin \theta \hat{\mathbf{p}}'_\Delta. \quad (2.212)$$

This vector is illustrated in Fig. 2.23.

Observe that the vector  $\mathbf{d}_{\Delta,\theta}$  is  $2N$ -dimensional, and within this space the orthogonal unit vectors  $\hat{\mathbf{p}}_\Delta$  and  $\hat{\mathbf{p}}'_\Delta$  define a two-dimensional subspace, which we will refer

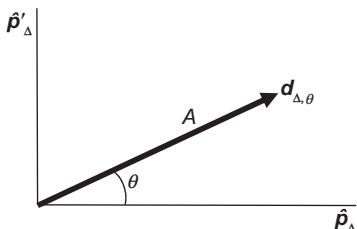


FIG. 2.23

Vector representation of  $N$  signal samples.

to as a “plane.” Thus the vector  $\mathbf{d}_{\Delta,\theta}$  resides in this two-dimensional subspace. Therefore each frequency offset  $\omega_{\Delta}$  is associated with a different two-dimensional plane. In the context of geometric algebra, one can say that *each frequency offset  $\omega_{\Delta}$  is associated with a different Doppler steering bivector*. The amplitude  $A$  of the signal samples is embodied in the magnitude of the vector  $\mathbf{d}_{\Delta,\theta}$  and the phase offset  $\theta$  is embodied in the amount by which  $\mathbf{d}_{\Delta,\theta}$  is rotated from the vector  $\hat{\mathbf{p}}_{\Delta}$  within this plane.

A more general sampled signal  $r(t_i)$ ,  $i = 1, \dots, N$  may be represented by a  $2N$ -dimensional real vector

$$\mathbf{d} = \begin{bmatrix} \mathcal{J}_1 \\ \vdots \\ \mathcal{J}_N \\ Q_1 \\ \vdots \\ Q_N \end{bmatrix} \quad (2.213)$$

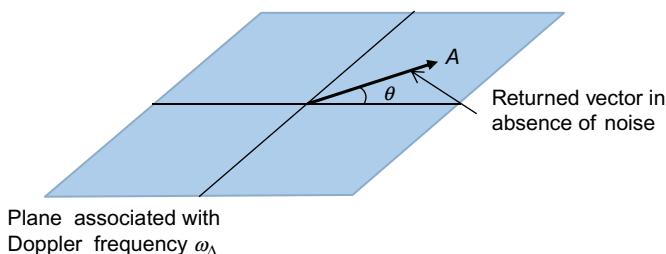
Here,  $\mathcal{J}_k$  and  $Q_k$  are the traditional in-phase and quadrature components of the  $k$ th sample. Traditionally these samples are gathered into an  $N$ -dimensional complex vector. Here they are represented by a  $2N$ -dimensional real vector. If this received signal  $r(t)$  has the form  $s(t) = A \cos((\omega_0 + \omega_{\Delta})t + \theta)$ , e.g., if the returned signal has Doppler frequency  $\omega_{\Delta}$ , amplitude  $A$ , and initial phase offset  $\theta$ , then as shown in Fig. 2.24, in the absence of noise the vector  $\mathbf{d} = \mathbf{d}_{\Delta,\theta}$  will lie in the  $\hat{\mathbf{p}}_{\Delta}, \hat{\mathbf{p}}'_{\Delta}$  plane and will rotate *within* this plane according to the value of the initial phase offset  $\theta$ .

If this received signal includes additive noise, then as shown in Fig. 2.25 the noise will rotate and dilate the vector  $\mathbf{d}_{\Delta,\theta}$  within the plane and cause it to move out of the plane. Any deviation outside of the two-dimensional subspace of the plane defined by  $\omega_{\Delta}$  is thus caused by noise.

#### 2.2.4.2 Geometric Interpretation of a Complex Data Vector as a Spinor Expansion

These ideas will now be used in conjunction with the notion of a spinor to give an alternate representation of a complex vector. To begin, consider a real vector in a  $2N$ -dimensional space and assume that it is confined to a two-dimensional subspace that has basis vectors  $\hat{\mathbf{e}}_1$  and  $\hat{\mathbf{e}}_{1+N}$ :

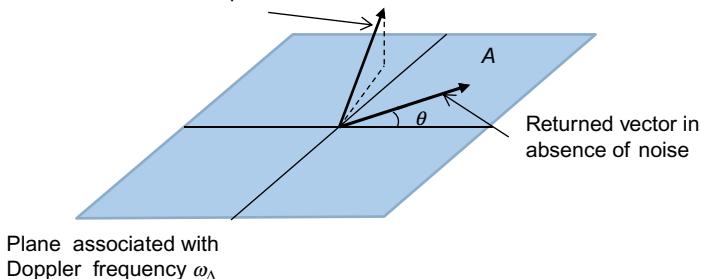
$$\mathbf{x} = x_1 \hat{\mathbf{e}}_1 + x_{1+N} \hat{\mathbf{e}}_{1+N} \quad (2.214)$$



**FIG. 2.24**

Returned signal vector in doppler steering plane.

Adding noise changes the length of the returned vector, rotates it within the plane, and moves it out of the plane



**FIG. 2.25**

Effect of noise on returned signal vector in doppler steering plane.

This is simply a vector in the plane defined by  $\hat{e}_1$  and  $\hat{e}_{1+N}$ . We already know that a complex number is represented in a plane so now use this vector to explore this notion further. In particular, factor out  $\hat{e}_1$  to the right to obtain

$$\mathbf{x} = (x_1 + x_{1+N}\hat{e}_{1+N}\hat{e}_1)\hat{e}_1 \quad (2.215)$$

Note that this factorization is permitted because  $\hat{e}_{1+N}\hat{e}_1\hat{e}_1 = \hat{e}_{1+N}$ . Now write

$$x_1 = x_1 + i_1 x_{1+N} \quad (2.216)$$

where  $i_1$  defines the unit bivector

$$i_1 = \hat{e}_{1+N}\hat{e}_1 \quad (2.217)$$

Thus the real two-dimensional vector  $\mathbf{x}$  may be written as

$$\mathbf{x} = c_1 \hat{e}_1 \quad (2.218)$$

where  $c_1$  is now a “complex number.” Now think of  $c_1 \hat{e}_1$  as the one-dimensional *complex* vector

$$c_1 \hat{e}_1 = [c_1] \quad (2.219)$$

Thus a one-dimensional complex vector may be viewed as two-dimensional vector confined to the plane defined by the relevant unit bivector that defines the complex number. As discussed earlier, the product of two vectors may be viewed as a spinor. In this example, multiplying  $\mathbf{x}$  by  $\hat{e}_1$  leads to

$$\mathbf{x}\hat{e}_1 = c_1 \hat{e}_1 \hat{e}_1 = c_1 \quad (2.220)$$

Thus one may view the complex vector

$$[c_1] = c_1 \hat{e}_1 = \mathbf{x} \quad (2.221)$$

as comprising the explicitly given spinor  $c_1$  that operates in the  $\hat{e}_1, \hat{e}_2$  plane to rotate and dilate the real unit vector  $\hat{e}_1$  into the real vector  $\mathbf{x}$  in that plane.

These ideas can be extended to more general complex vectors. Begin with an arbitrary real vector  $\mathbf{d}$  in  $R^{2N}$  and assume an orthonormal basis  $\{\hat{\mathbf{e}}_l, l=1, \dots, 2N\}$  for this space. With this configuration one may write the vector  $\mathbf{d}$  as

$$\mathbf{d} = \sum_{k=1}^N \mathbf{d}_k \quad (2.222)$$

with

$$\mathbf{d} = d_{1,k} \hat{\mathbf{e}}_k + d_{2,k} \hat{\mathbf{e}}_{k+N} \quad (2.223)$$

$$d_{1,k} = \mathbf{d} \cdot \hat{\mathbf{e}}_k \quad (2.224)$$

$$d_{2,k} = \mathbf{d} \cdot \hat{\mathbf{e}}_{k+N}. \quad (2.225)$$

This is simply a straightforward expansion of  $\mathbf{d}$  by projecting it into  $N$  orthogonal two-dimensional subspaces. Now use the geometric product and  $\hat{\mathbf{e}}_k \hat{\mathbf{e}}_k = 1$  to define

$$c_k = \mathbf{d}_k \hat{\mathbf{e}}_k = d_{1,k} + i_k d_{2,k} \quad (2.226)$$

where

$$i_k = \hat{\mathbf{e}}_{k+N} \hat{\mathbf{e}}_k \quad (2.227)$$

is a unit bivector in the  $\hat{\mathbf{e}}_k, \hat{\mathbf{e}}_{k+N}$  plane. The quantity  $c_k$  is spinor and as discussed before may be thought of as being analogous to a complex number. However, it is restricted to the  $\hat{\mathbf{e}}_{k+N}, \hat{\mathbf{e}}_k$  plane where the bivector  $i_k$  replaces  $i = \sqrt{-1}$ . With this definition multiplying both sides of Eq. (2.226) by  $\hat{\mathbf{e}}_k$  leads to

$$\mathbf{d}_k = c_k \hat{\mathbf{e}}_k. \quad (2.228)$$

Again, the spinor  $c_k$  may be viewed as an operator that transforms the real unit vector  $\hat{\mathbf{e}}_k$  into the real vector  $\mathbf{d}_k$  in the  $\hat{\mathbf{e}}_k, \hat{\mathbf{e}}_{k+N}$  plane. Now use Eq. (2.228) in Eq. (2.222) to write

$$\mathbf{d} = \sum_{k=1}^N c_k \hat{\mathbf{e}}_k. \quad (2.229)$$

As discussed before, the spinor  $c_k$  may be obtained by “projecting” (*not an orthogonal projection*) the full vector  $\mathbf{d}$  onto the bivector  $i_k$  as follows:

$$c_k = -(\mathbf{d} \cdot i_k) \hat{\mathbf{e}}_{k+N}. \quad (2.230)$$

This result may also be written as

$$\mathbf{d} = \sum_{k=1}^N (\mathbf{d} \cdot i_k) i_k^\dagger \quad (2.231)$$

where

$$i_k^\dagger = -\hat{\mathbf{e}}_{k+N} \hat{\mathbf{e}}_k = \hat{\mathbf{e}}_k \hat{\mathbf{e}}_{k+N} \quad (2.232)$$

This formulation represents the  $2N$ -dimensional vector  $\mathbf{d}$  as an  $N$ -dimensional complex vector with a definite geometric interpretation. To explore this assertion, first note that

$$|\mathbf{d}|^2 = \sum_{k=1}^N |\mathbf{d}_k|^2 \quad (2.233)$$

and

$$|\mathbf{d}_k|^2 = d_{1,k}^2 + d_{2,k}^2 = |c_k|^2. \quad (2.234)$$

Thus

$$|\mathbf{d}|^2 = \sum_{k=1}^N |c_k|^2. \quad (2.235)$$

This relation holds for any choice of the orthonormal basis  $\{\hat{\mathbf{e}}_l, l = 1, \dots, 2N\}$ . In addition, it is also straightforward to show

$$|\mathbf{d}_k| = |\mathbf{d}| \cos \psi_k \quad (2.236)$$

for some angle  $\psi_k$ . Finally, using the fact that  $i_k^2 = -1$ , one may rewrite  $c_k$  as

$$c_k = |c_k| e^{i_k \theta_k} = |\mathbf{d}| \cos \psi_k e^{i_k \theta_k}. \quad (2.237)$$

These relationships are illustrated in Fig. 2.26.

Because  $c_k, k = 1, \dots, N$  may be interpreted as complex numbers in their respective two-dimensional subspaces (i.e., planes), the complex vector expansion of the vector  $\mathbf{d}$  in Eq. (2.229) is analogous to a real vector expansion. Instead of projecting a vector into  $N$  one-dimensional subspaces represented by unit vectors  $\hat{\mathbf{e}}_k$ , the formulation projects  $\mathbf{d}$  into  $N$  two-dimensional subspaces represented by the unit bivectors  $\hat{\mathbf{e}}_{k+N} \hat{\mathbf{e}}_k$ . Moreover, the *inner product* of  $\mathbf{d}$  with the unit bivector  $\hat{\mathbf{e}}_{k+N} \hat{\mathbf{e}}_k$  leads to the spinor  $c_k$ . If  $\mathbf{d}$  were instead projected into the one-dimensional subspaces  $\hat{\mathbf{e}}_k$ , the appropriate scalar would be obtained from the inner product of  $\mathbf{d}$  with the corresponding unit vector  $\hat{\mathbf{e}}_k$ . Finally, the operation of the spinor  $c_k$  on the unit vector  $\hat{\mathbf{e}}_k$  yields the projected vector  $\mathbf{d}_k$ . This is similar to the operation of a scalar on a unit vector to obtain the corresponding projected vector.

To further understand the spinor expansion before, consider a specific orthonormal basis. We will reserve the notation  $\hat{\mathbf{e}}_l$  and  $c_l$  for general unspecified bases and their corresponding spinors and will assign different notations for specific choices of the basis and spinor. In particular, let

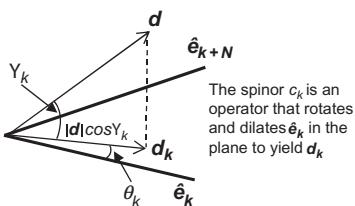


FIG. 2.26

Spinor interpretation.

$$\hat{\varphi}_l = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad l = 1, \dots, 2N \quad (2.238)$$

where the nonzero entry is in the  $l$ th position. We call this the *observed data basis*. It is clearly orthonormal. If we have a sequence of  $N$  observed data samples, for example in-phase and quadrature samples  $\mathcal{J}_k, \mathcal{Q}_k, k = 1, \dots, N$ , then in the observed data basis we use  $i_{k,\varphi} = \hat{\varphi}_{k+N} \hat{\varphi}_k$  and set

$$\mathbf{d} = \begin{bmatrix} \mathcal{J}_1 \\ \vdots \\ \mathcal{J}_N \\ \mathcal{Q}_1 \\ \vdots \\ \mathcal{Q}_N \end{bmatrix} \quad (2.239)$$

$$f_k = \mathcal{J}_k + i_{k,\varphi} \mathcal{Q}_k, \quad k = 1, \dots, N \quad (2.240)$$

The expansion becomes

$$\mathbf{d} = \sum_{k=1}^N f_k \hat{\varphi}_k. \quad (2.241)$$

We may interpret the right-hand side of Eq. (2.241) as giving rise to the complex vector

$$\mathbf{f} = \begin{bmatrix} \mathcal{J}_1 + i_{1,\varphi} \mathcal{Q}_1 \\ \vdots \\ \mathcal{J}_N + i_{N,\varphi} \mathcal{Q}_N \end{bmatrix}. \quad (2.242)$$

In this interpretation, each observed complex sample  $f_k = \mathcal{J}_k + i_{k,\varphi} \mathcal{Q}_k$  is a spinor that operates in its respective  $\hat{\varphi}_{k+N}, \hat{\varphi}_k$  plane to transform the unit vector  $\hat{\varphi}_k$  into the real vector

$$\begin{aligned} \mathbf{d}_k &= \operatorname{Re}(f_k) \hat{\varphi}_k + \operatorname{Im}(f_k) \hat{\varphi}_{k+N} \\ &= \mathcal{J}_k \hat{\varphi}_k + \mathcal{Q}_k \hat{\varphi}_{k+N} \end{aligned} \quad (2.243)$$

which also lies in the  $\hat{\varphi}_{k+N}, \hat{\varphi}_k$  plane. Thus we may interpret  $\mathcal{J}$  and  $\mathcal{Q}$  data, which traditionally is represented as a complex vector, as either a  $2N$ -dimensional real vector or a collection of  $N$  operators (spinors) that operate on unit vectors in  $N$  two-dimensional subspaces of a  $2N$ -dimensional space.

### 2.2.4.3 Projecting a Vector into a Subspace

As is well known, a vector can be projected onto another vector, for example, a unit vector that forms part of a basis spanning a space. The unit vectors represent one-dimensional subspaces of the space in which the vector is found. In addition, as

was discussed before, a complex vector may be thought of a real vector of a twice the dimension whose components are obtained by projecting the given vector onto unit bivectors that span the space. The unit bivectors represent two-dimensional subspaces of the space in which the vector is found. Of course, it is possible to project a vector into any dimension subspace. This section explores how to project a vector onto more general subspaces in the context of geometric algebra.

Let  $\mathbf{r}$  be an  $N$ -dimensional real vector. Let  $A_N = \lambda_N I_N$  where  $I_N$  is the pseudoscalar for the associated geometric algebra and  $\lambda_N$  is a scalar. Let

$$A_N = A_{k_1} \wedge A_{k_2} \wedge \cdots \wedge A_{k_m} \quad (2.244)$$

where  $A_{k_j}, j = 1, \dots, m, 1 \leq m \leq N$  is a  $k_j$ -blade associated with a  $k_j$ -dimensional subspace of  $R^N$  where the subspaces are distinct with  $k_1 + k_2 + \cdots + k_m = N$  (hence  $A_{k_j}$  is proportional to  $I_{k_j}$  where  $I_{k_j}$  is the pseudoscalar of the given subspace). Since the space is  $N$ -dimensional, let  $\{\mathbf{a}_1, \dots, \mathbf{a}_N\}$  be a set linearly independent vectors that span the space. Then one has

$$A_{k_1} = \mathbf{a}_1 \wedge \mathbf{a}_2 \wedge \cdots \wedge \mathbf{a}_{k_1} \quad (2.245)$$

$$A_{k_2} = \mathbf{a}_{k_1+1} \wedge \cdots \wedge \mathbf{a}_{k_1+k_2} \quad (2.246)$$

⋮

$$A_{k_m} = \mathbf{a}_{k_{m-1}+1} \wedge \cdots \wedge \mathbf{a}_{k_{m-1}+k_m} \quad (2.247)$$

In effect this formulation decomposes  $R^N$  into  $m$  distinct subspaces where the  $i$ th subspace has size  $k_i, i = 1, \dots, m$  and thus the subspaces are not necessarily of the same size. For example, if  $m = N$  then  $k_1 = k_2 = \cdots = k_m = 1$  and the  $N$  subspaces are  $N$  linearly independent vectors. On the other hand, if  $m = 2$ , then there are two distinct subspaces of sizes  $k_1$  and  $k_2 = N - k_1$ . There is great flexibility in how  $R^N$  may be decomposed into subspaces.

With this formulation write

$$\mathbf{r} = \mathbf{r} A_{k_1} A_{k_1}^{-1} \quad (2.248)$$

Note that this is simply an identity since  $A_{k_1} A_{k_1}^{-1} = 1$ . The geometric product of a vector and a blade now leads to

$$\begin{aligned} \mathbf{r} &= (\mathbf{r} \cdot A_{k_1} + \mathbf{r} \wedge A_{k_1}) A_{k_1}^{-1} \\ &= \mathbf{r}_{k_1} + \mathbf{r}_{k_1, \perp} \end{aligned} \quad (2.249)$$

where

$$\mathbf{r}_{k_1} = (\mathbf{r} \cdot A_{k_1}) A_{k_1}^{-1} \text{ is the } \textit{projection} \text{ of } \mathbf{r} \text{ into the subspace } I_{k_1} \quad (2.250)$$

$$\mathbf{r}_{k_1, \perp} = (\mathbf{r} \wedge A_{k_1}) A_{k_1}^{-1} \text{ is the } \textit{rejection} \text{ of } \mathbf{r} \text{ from the subspace } I_{k_1} \quad (2.251)$$

This decomposes  $\mathbf{r}$  into a component  $\mathbf{r}_{k_1}$  that is in the subspace  $I_{k_1}$  and a component  $\mathbf{r}_{k_1, \perp}$  that is not in this subspace. Repeating this process with the vector  $\mathbf{r}_{k_1, \perp}$  now

projects  $\mathbf{r}_{k_1,\perp}$  into a component in the subspace  $I_{k_2}$  and a component not in  $I_{k_2}$ . Ultimately this process yields

$$\mathbf{r} = \sum_{j=1}^m (\mathbf{r} \cdot A_{k_j}) A_{k_j}^{-1} = (\mathbf{r} \cdot A_{k_1}) A_{k_1}^{-1} + (\mathbf{r} \cdot A_{k_2}) A_{k_2}^{-1} + \cdots + (\mathbf{r} \cdot A_{k_m}) A_{k_m}^{-1} \quad (2.252)$$

This presents a decomposition of the vector  $\mathbf{r}$  in terms of its projection into  $m$  distinct subspaces that span  $R^N$ . Using the earlier result about the inverse of a nonzero blade, this result may finally be written as

$$\mathbf{r} = \sum_{j=1}^m (\mathbf{r} \cdot A_{k_j}) \frac{A_{k_j}^\dagger}{|A_{k_j}|^2} \quad (2.253)$$

#### 2.2.4.3.1 Examples

1. Consider first the case  $m=N$ . In this case the subspaces  $A_{k_j}$  are simply one-dimensional vectors, and thus

$$A_{k_j} = \mathbf{a}_j, \quad j = 1, \dots, N \quad (2.254)$$

The inverse of a vector now yields

$$A_{k_j}^{-1} = \frac{A_{k_j}^\dagger}{|A_{k_j}|^2} = \frac{\mathbf{a}_j}{|\mathbf{a}_j|^2}, \quad j = 1, \dots, N \quad (2.255)$$

Thus the expansion becomes

$$\mathbf{r} = \sum_{j=1}^m (\mathbf{r} \cdot A_{k_j}) \frac{A_{k_j}^\dagger}{|A_{k_j}|^2} = (\mathbf{r} \cdot \mathbf{a}_1) \frac{\mathbf{a}_1}{|\mathbf{a}_1|^2} + (\mathbf{r} \cdot \mathbf{a}_2) \frac{\mathbf{a}_2}{|\mathbf{a}_2|^2} + \cdots + (\mathbf{r} \cdot \mathbf{a}_N) \frac{\mathbf{a}_N}{|\mathbf{a}_N|^2} \quad (2.256)$$

which is the usual expansion of a vector in terms of a set of linearly independent vectors  $\mathbf{a}_j, j = 1, \dots, N$  that span the space.

2. As another example, assume  $N=2M$  is even and let  $m=M$ . Hence the task is to decompose  $R^N$  into  $M$  two-dimensional subspaces. Write

$$A_{k_1} = \mathbf{a}_{M+1} \wedge \mathbf{a}_1 \quad (2.257)$$

$$A_{k_2} = \mathbf{a}_{M+2} \wedge \mathbf{a}_2 \quad (2.258)$$

⋮

$$A_{k_M} = \mathbf{a}_{2M} \wedge \mathbf{a}_M \quad (2.259)$$

Thus  $A_{k_j}, j = 1, \dots, M$  are bivectors that span the space and define the two-dimensional subspaces. From the inner product of a vector with a bivector:

$$\mathbf{r} \cdot A_{k_j} = \mathbf{r} \cdot (\mathbf{a}_{M+j} \wedge \mathbf{a}_j) = (\mathbf{r} \cdot \mathbf{a}_{M+j}) \mathbf{a}_j - (\mathbf{r} \cdot \mathbf{a}_j) \mathbf{a}_{M+j} \quad (2.260)$$

Let  $\hat{\mathbf{e}}_j$  and  $\hat{\mathbf{e}}_{M+j}$  be orthogonal unit vectors in the plane of  $\mathbf{a}_j$  and  $\mathbf{a}_{M+j}$  and write

$$\mathbf{a}_j = (\mathbf{a}_j \cdot \hat{\mathbf{e}}_j) \hat{\mathbf{e}}_j + (\mathbf{a}_j \cdot \hat{\mathbf{e}}_{M+j}) \hat{\mathbf{e}}_{M+j} \quad (2.261)$$

$$\mathbf{a}_{M+j} = (\mathbf{a}_{M+j} \cdot \hat{\mathbf{e}}_j) \hat{\mathbf{e}}_j + (\mathbf{a}_{M+j} \cdot \hat{\mathbf{e}}_{M+j}) \hat{\mathbf{e}}_{M+j} \quad (2.262)$$

It should be clear that ultimately the bivector  $\mathbf{a}_j \wedge \mathbf{a}_{M+j}$  is a weighted version of the unit bivector  $\hat{\mathbf{e}}_{M+j} \hat{\mathbf{e}}_j$  defined in the plane of  $\mathbf{a}_j$  and  $\mathbf{a}_{M+j}$ . To see this relationship explicitly, examine

$$\begin{aligned} \mathbf{a}_{M+j} \wedge \mathbf{a}_j &= [(\mathbf{a}_{M+j} \cdot \hat{\mathbf{e}}_j) \hat{\mathbf{e}}_j + (\mathbf{a}_{M+j} \cdot \hat{\mathbf{e}}_{M+j}) \hat{\mathbf{e}}_{M+j}] \wedge [(\mathbf{a}_j \cdot \hat{\mathbf{e}}_j) \hat{\mathbf{e}}_j + (\mathbf{a}_j \cdot \hat{\mathbf{e}}_{M+j}) \hat{\mathbf{e}}_{M+j}] \\ &= [(\mathbf{a}_{M+j} \cdot \hat{\mathbf{e}}_{M+j}) (\mathbf{a}_j \cdot \hat{\mathbf{e}}_j) - (\mathbf{a}_{M+j} \cdot \hat{\mathbf{e}}_j) (\mathbf{a}_j \cdot \hat{\mathbf{e}}_{M+j})] \hat{\mathbf{e}}_{M+j} \hat{\mathbf{e}}_j \\ &= |\mathbf{a}_{M+j} \wedge \mathbf{a}_j| \hat{\mathbf{e}}_{M+j} \hat{\mathbf{e}}_j \\ &= |\mathbf{a}_{M+j} \wedge \mathbf{a}_j| i_{k_j} \end{aligned} \quad (2.263)$$

where

$$\begin{aligned} |\mathbf{a}_{M+j} \wedge \mathbf{a}_j| &= (\mathbf{a}_{M+j} \cdot \hat{\mathbf{e}}_{M+j}) (\mathbf{a}_j \cdot \hat{\mathbf{e}}_j) - (\mathbf{a}_{M+j} \cdot \hat{\mathbf{e}}_j) (\mathbf{a}_j \cdot \hat{\mathbf{e}}_{M+j}) \\ i_{k_j} &= \hat{\mathbf{e}}_{M+j} \hat{\mathbf{e}}_j \end{aligned} \quad (2.264)$$

Following the same steps as before shows

$$A_{k_j}^\dagger = \mathbf{a}_j \wedge \mathbf{a}_{M+j} = |\mathbf{a}_j \wedge \mathbf{a}_{M+j}| \hat{\mathbf{e}}_j \hat{\mathbf{e}}_{M+j} = -|\mathbf{a}_{M+j} \wedge \mathbf{a}_j| i_{k_j}^\dagger \quad (2.265)$$

with

$$|\mathbf{a}_j \wedge \mathbf{a}_{M+j}| = -|\mathbf{a}_{M+j} \wedge \mathbf{a}_j| \quad (2.266)$$

$$i_{k_j}^\dagger = \hat{\mathbf{e}}_j \hat{\mathbf{e}}_{M+j} \quad (2.267)$$

Thus

$$|A_{k_j}|^2 = A_{k_j} A_{k_j}^\dagger = |\mathbf{a}_{M+j} \wedge \mathbf{a}_j|^2 \quad (2.268)$$

and

$$\frac{A_{k_j}^\dagger}{|A_{k_j}|^2} = \frac{i_{k_j}^\dagger}{|\mathbf{a}_{M+j} \wedge \mathbf{a}_j|} \quad (2.269)$$

Therefore in this example

$$(\mathbf{r} \cdot A_{k_j}) \frac{A_{k_j}^\dagger}{|A_{k_j}|^2} = (\mathbf{r} \cdot |\mathbf{a}_{M+j} \wedge \mathbf{a}_j| i_{k_j}) \frac{i_{k_j}^\dagger}{|\mathbf{a}_{M+j} \wedge \mathbf{a}_j|} \quad (2.270)$$

Because  $|\mathbf{a}_{M+j} \wedge \mathbf{a}_j|$  is simply a scalar, this expression reduces to

$$(\mathbf{r} \cdot A_{k_j}) \frac{A_{k_j}^\dagger}{|A_{k_j}|^2} = (\mathbf{r} \cdot i_{k_j}) i_{k_j}^\dagger \quad (2.271)$$

Thus in this example the expansion into two-dimensional subspaces takes the form

$$\mathbf{r} = \sum_{j=1}^M (\mathbf{r} \cdot i_{k_j}) i_{k_j}^\dagger \quad (2.272)$$

This is the same result that was obtained before in the discussion about representing complex vectors as real vectors in a space of twice the dimension. Hence this is equivalent to an expansion of the vector  $\mathbf{r}$  in terms of the two-dimensional subspaces of  $R^N$  defined by the bivectors  $A_{k_j}, j=1, \dots, M$  where  $N=2M$ :

$$\mathbf{r} = \sum_{j=1}^M (\mathbf{r} \cdot A_{k_j}) \frac{A_{k_j}^\dagger}{|A_{k_j}|^2} = \sum_{j=1}^M \{ (\mathbf{r} \cdot \hat{\mathbf{e}}_j) \hat{\mathbf{e}}_j + (\mathbf{r} \cdot \hat{\mathbf{e}}_{M+j}) \hat{\mathbf{e}}_{M+j} \} \quad (2.273)$$

Now define

$$c_j = (\mathbf{r} \cdot A_{k_j}) \frac{A_{k_j}^\dagger}{|A_{k_j}|^2} \hat{\mathbf{e}}_j = (\mathbf{r} \cdot \mathbf{e}_j) + i_{k_j} (\mathbf{r} \cdot \hat{\mathbf{e}}_{M+j}) \quad (2.274)$$

The earlier discussion showed that  $c_j$  is the geometric algebra representation of a complex number and represents a spinor in the plane of  $\mathbf{a}_j$  and  $\mathbf{a}_{M+j}$ . Hence

$$\mathbf{r} = \sum_{j=1}^M c_j \hat{\mathbf{e}}_j \quad (2.275)$$

This relation represents the  $2M$ -dimensional vector  $\mathbf{r}$  as the  $M$ -dimensional “complex vector”

$$\begin{bmatrix} c_1 \\ \vdots \\ c_M \end{bmatrix} \quad (2.276)$$

Note that one has great latitude in choosing the orthonormal vectors  $\mathbf{e}_j$  and  $\hat{\mathbf{e}}_{M+j}$  within the plane of  $\mathbf{a}_j$  and  $\mathbf{a}_{M+j}$ .

To summarize, the  $N$ -dimensional space  $R^N$  may be decomposed into  $m$  distinct subspaces where the sizes of the subspaces are  $k_j, j=1, \dots, m$ ,  $1 \leq m \leq N$  with  $k_1 + k_2 + \dots + k_m = N$  (so the subspaces are not necessarily of the same size) and a vector  $\mathbf{r}$  may be projected into these subspaces as

$$\mathbf{r} = \sum_{j=1}^m (\mathbf{r} \cdot A_{k_j}) \frac{A_{k_j}^\dagger}{|A_{k_j}|^2} \quad (2.277)$$

where  $A_{k_j}, j=1, \dots, m$  are  $k_j$ -blades defining the relevant subspaces. Observe that for vectors,  $\mathbf{a}^\dagger = \mathbf{a}$ ; hence this formulation is a direct generalization of the well-known approach to expanding  $\mathbf{r}$  in one-dimensional subspaces (i.e., vectors).

In the previous example, expanding  $\mathbf{r}$  by projecting it into two-dimensional subspaces led to an expression of the form

$$\mathbf{r} = \sum_{j=1}^m (\mathbf{r} \cdot i_{k_j}) i_{k_j}^\dagger \quad (2.278)$$

where  $\{i_{k_j}, j=1, \dots, m\}$  were mutually orthogonal unit bivectors. It turns out to be generally true that the blade  $A_{k_j}$  is proportional to the pseudoscalar  $I_{k_j}$  associated with the subspace corresponding to  $A_{k_j}$ . As result, the general result before may be written as

$$\mathbf{r} = \sum_{j=1}^m (\mathbf{r} \cdot I_{k_j}) I_{k_j}^\dagger \quad (2.279)$$

The formulation in Eq. (2.277) represents each distinct subspace by a set of linearly independent vectors  $\{\mathbf{a}_j\}$  whereas the formulation in Eq. (2.279) represents each subspace by a set of *orthonormal* vectors that may be used to form the relevant pseudoscalar  $I_{k_j}$ . This is fully analogous to the well-known expansion of a vector by projecting it onto one-dimensional vectors.

### 2.2.5 WHAT IS A COMPLEX MATRIX?

Let  $\mathbf{c}_{complex}$  be an arbitrary  $N$ -dimensional complex vector:

$$\mathbf{c}_{complex} = \mathbf{c}_r + i\mathbf{c}_i \quad (2.280)$$

As discussed in the earlier sections, in general we associate a  $2N$ -dimensional vector  $\mathbf{c}$  with the complex vector  $\mathbf{c}_{complex}$  via:

$$\mathbf{c}_{complex} \rightarrow \mathbf{c} = \begin{bmatrix} \mathbf{c}_r \\ \mathbf{c}_i \end{bmatrix} \quad (2.281)$$

Now assume  $\boldsymbol{\phi}$  and  $\boldsymbol{\phi}'$  are  $2N$ -dimensional unit vectors associated with  $N$ -dimensional complex vectors  $\boldsymbol{\phi}_{complex} = \boldsymbol{\phi}_r + i\boldsymbol{\phi}_i$  and  $\boldsymbol{\phi}_{complex}' = -\boldsymbol{\phi}_i + i\boldsymbol{\phi}_r$  by

$$\boldsymbol{\phi}_{complex} \rightarrow \boldsymbol{\phi} = \begin{bmatrix} \boldsymbol{\phi}_r \\ \boldsymbol{\phi}_i \end{bmatrix} \quad (2.282)$$

$$\boldsymbol{\phi}'_{complex} \rightarrow \boldsymbol{\phi}' = \begin{bmatrix} -\boldsymbol{\phi}_i \\ \boldsymbol{\phi}_r \end{bmatrix} \quad (2.283)$$

Note that the two complex vectors are “orthogonal” to each other in the sense that

$$\boldsymbol{\phi}'_{complex} = i(\boldsymbol{\phi}_{complex}) \quad (2.284)$$

For the corresponding  $2N$ -dimensional vectors it follows that

$$\boldsymbol{\phi} \cdot \boldsymbol{\phi}' = \begin{bmatrix} \boldsymbol{\phi}_r \\ \boldsymbol{\phi}_i \end{bmatrix} \cdot \begin{bmatrix} -\boldsymbol{\phi}_i \\ \boldsymbol{\phi}_r \end{bmatrix} = -\boldsymbol{\phi}_r \cdot \boldsymbol{\phi}_i + \boldsymbol{\phi}_i \cdot \boldsymbol{\phi}_r = 0 \quad (2.285)$$

Thus  $\boldsymbol{\phi}$  and  $\boldsymbol{\phi}'$  are orthonormal and form a basis for a plane in the  $2N$ -dimensional space. Now let  $\lambda_{complex} = \lambda_r + i\lambda_i$  be a complex scalar and examine

$$\begin{aligned} \lambda_{complex} \boldsymbol{\phi}_{complex} &= (\lambda_r + i\lambda_i)(\boldsymbol{\phi}_r + i\boldsymbol{\phi}_i) \\ &= \lambda_r \boldsymbol{\phi}_r - \lambda_i \boldsymbol{\phi}_i + i(\lambda_r \boldsymbol{\phi}_i + \lambda_i \boldsymbol{\phi}_r) \end{aligned} \quad (2.286)$$

From the correspondence defined before:

$$\begin{aligned}\lambda_r \boldsymbol{\phi}_r - \lambda_i \boldsymbol{\phi}_i + i(\lambda_r \boldsymbol{\phi}_i + \lambda_i \boldsymbol{\phi}_r) &\rightarrow \begin{bmatrix} \lambda_r \boldsymbol{\phi}_r - \lambda_i \boldsymbol{\phi}_i \\ \lambda_r \boldsymbol{\phi}_i + \lambda_i \boldsymbol{\phi}_r \end{bmatrix} \\ &= \lambda_r \begin{bmatrix} \boldsymbol{\phi}_r \\ \boldsymbol{\phi}_i \end{bmatrix} + \lambda_i \begin{bmatrix} -\boldsymbol{\phi}_i \\ \boldsymbol{\phi}_r \end{bmatrix} \\ &= \lambda_r \boldsymbol{\phi} + \lambda_i \boldsymbol{\phi}'\end{aligned}\quad (2.287)$$

Thus the multiplication of an  $N$ -dimensional complex vector by a complex scalar leads to the sum of two real  $2N$ -dimensional vectors given by the following correspondence:

$$\lambda_{\text{complex}} \boldsymbol{\phi}_{\text{complex}} \rightarrow \lambda_r \boldsymbol{\phi} + \lambda_i \boldsymbol{\phi}' \quad (2.288)$$

This result shows that the effect of multiplying the  $N$ -dimensional complex vector  $\boldsymbol{\phi}_{\text{complex}}$  by the complex scalar  $\lambda_{\text{complex}}$  is to rotate and dilate the corresponding  $2N$ -dimensional vector in the plane shared by  $\boldsymbol{\phi}$  and  $\boldsymbol{\phi}'$  as illustrated in Fig. 2.27.

Similarly

$$\begin{aligned}\lambda_{\text{complex}} \boldsymbol{\phi}'_{\text{complex}} &= (\lambda_r + i\lambda_i)(-\boldsymbol{\phi}_i + i\boldsymbol{\phi}_r) \\ &= -\lambda_r \boldsymbol{\phi}_i - \lambda_i \boldsymbol{\phi}_r + i(-\lambda_i \boldsymbol{\phi}_i + \lambda_r \boldsymbol{\phi}_r) \\ &\rightarrow \begin{bmatrix} -\lambda_r \boldsymbol{\phi}_i - \lambda_i \boldsymbol{\phi}_r \\ -\lambda_i \boldsymbol{\phi}_i + \lambda_r \boldsymbol{\phi}_r \end{bmatrix} \\ &= \lambda_r \begin{bmatrix} -\boldsymbol{\phi}_i \\ \boldsymbol{\phi}_r \end{bmatrix} - \lambda_i \begin{bmatrix} \boldsymbol{\phi}_r \\ \boldsymbol{\phi}_i \end{bmatrix} \\ &= \lambda_r \boldsymbol{\phi}' - \lambda_i \boldsymbol{\phi}\end{aligned}\quad (2.289)$$

As illustrated in Fig. 2.28, multiplying the complex vector  $\boldsymbol{\phi}'_{\text{complex}}$  by the complex scalar  $\lambda_{\text{complex}}$  similarly rotates and dilates the corresponding  $2N$ -dimensional vector in the plane shared by  $\boldsymbol{\phi}$  and  $\boldsymbol{\phi}'$ .

Therefore any vector expressed in the  $\boldsymbol{\phi}, \boldsymbol{\phi}'$  basis—i.e., in the plane defined by  $\boldsymbol{\phi}$  and  $\boldsymbol{\phi}'$ —is dilated and rotated (in the same plane) by a spinor  $\lambda_s$  obtained from the complex scalar  $\lambda_{\text{complex}}$  via

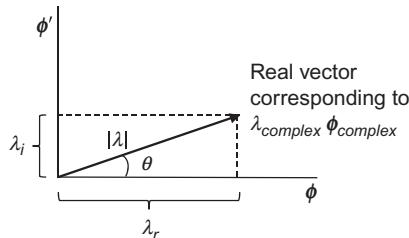
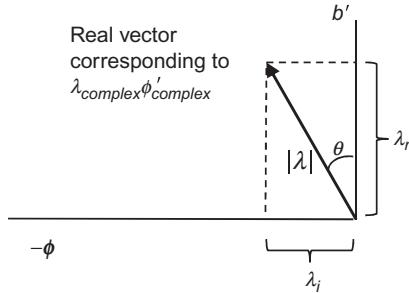


FIG. 2.27

Rotation of a complex vector by a complex scalar.

**FIG. 2.28**

Rotation of a complex vector by a complex scalar.

$$\lambda_s = \lambda_r + i_{\phi'\phi}\lambda_i \quad (2.290)$$

where

$$i_{\phi'\phi} = \phi' \wedge \phi = \phi' \phi \quad (2.291)$$

defines a unit bivector in the  $\phi, \phi'$  plane. Note that this unit bivector is defined so as to rotate  $\phi$  into  $\phi'$  via left-multiplication:

$$i_{\phi'\phi}\phi = \phi' \phi \phi = \phi' \quad (2.292)$$

Similarly

$$i_{\phi'\phi}\phi' = \phi' \phi \phi' = -\phi \quad (2.293)$$

With this notation

$$\begin{aligned} \lambda_s \phi &= (\lambda_r + i_{\phi'\phi}\lambda_i) \phi \\ &= \lambda_r \phi + \lambda_i i_{\phi'\phi} \phi \\ &= \lambda_r \phi + \lambda_i \phi' \end{aligned} \quad (2.294)$$

$$\begin{aligned} \lambda_s \phi' &= (\lambda_r + i_{\phi'\phi}\lambda_i) \phi' \\ &= \lambda_r \phi' + \lambda_i i_{\phi'\phi} \phi' \\ &= \lambda_r \phi' - \lambda_i \phi \end{aligned} \quad (2.295)$$

Hence

$$\lambda_{complex}\phi_{complex} \rightarrow \lambda_s \phi \quad (2.296)$$

$$\lambda_{complex}\phi'_{complex} \rightarrow \lambda_s \phi' \quad (2.297)$$

From earlier work we know that the action of a spinor on a vector in the same plane is to dilate and rotate it in the plane. To see the rotation and dilation effect more clearly, let

$$\lambda_r = |\lambda| \cos \theta \quad (2.298)$$

$$\lambda_i = |\lambda| \sin \theta \quad (2.299)$$

Represent the spinor  $\lambda_s$ , which operates in the  $\phi, \phi'$  plane, as

$$\lambda_s = |\lambda| \cos \theta + i_{\phi'\phi} |\lambda| \sin \theta = |\lambda| e^{i_{\phi'\phi}\theta} \quad (2.300)$$

Let  $\mathbf{x}$  be an arbitrary vector in the  $\phi, \phi'$  plane:

$$\mathbf{x} = (\mathbf{x} \cdot \phi) \phi + (\mathbf{x} \cdot \phi') \phi' \quad (2.301)$$

Then

$$\begin{aligned} \lambda_s \mathbf{x} &= (|\lambda| \cos \theta + i_{\phi' \phi} |\lambda| \sin \theta) ((\mathbf{x} \cdot \phi) \phi + (\mathbf{x} \cdot \phi') \phi') \\ &= |\lambda| \cos \theta ((\mathbf{x} \cdot \phi) \phi + (\mathbf{x} \cdot \phi') \phi') + i_{\phi' \phi} |\lambda| \sin \theta ((\mathbf{x} \cdot \phi) \phi + (\mathbf{x} \cdot \phi') \phi') \\ &= |\lambda| \cos \theta (\mathbf{x} \cdot \phi) \phi + |\lambda| \cos \theta (\mathbf{x} \cdot \phi') \phi' + |\lambda| \sin \theta (\mathbf{x} \cdot \phi) \phi' - |\lambda| \sin \theta (\mathbf{x} \cdot \phi') \phi \\ &= |\lambda| (\cos \theta (\mathbf{x} \cdot \phi) - \sin \theta (\mathbf{x} \cdot \phi')) \phi + |\lambda| (\cos \theta (\mathbf{x} \cdot \phi') + \sin \theta (\mathbf{x} \cdot \phi)) \phi' \end{aligned} \quad (2.302)$$

Now let

$$\mathbf{x} \cdot \phi = |\mathbf{x}| \cos \theta_x \quad (2.303)$$

$$\mathbf{x} \cdot \phi' = |\mathbf{x}| \sin \theta_x \quad (2.304)$$

Substitution yields

$$\begin{aligned} \lambda_s \mathbf{x} &= |\lambda| |\mathbf{x}| (\cos \theta \cos \theta_x - \sin \theta \sin \theta_x) \phi + |\lambda| |\mathbf{x}| (\cos \theta \sin \theta_x + \sin \theta \cos \theta_x) \phi' \\ &= |\lambda| |\mathbf{x}| \cos(\theta + \theta_x) \phi + |\lambda| |\mathbf{x}| \sin(\theta + \theta_x) \phi' \end{aligned} \quad (2.305)$$

Thus the spinor  $\lambda_s$  rotates and dilates the vector  $\mathbf{x}$  in the  $\phi, \phi'$  plane as shown in Fig. 2.29:

Now consider a complex matrix  $L = L_r + iL_i$  with complex eigenvalue  $\lambda$  and corresponding complex eigenvector  $\phi_{complex}$ :

$$L \phi_{complex} = \lambda_{complex} \phi_{complex} \quad (2.306)$$

This relation may be rewritten as

$$L_r \phi_r - L_i \phi_i + i(L_r \phi_i + L_i \phi_r) = \lambda_r \phi_r - \lambda_i \phi_i + i(\lambda_r \phi_i + \lambda_i \phi_r) \quad (2.307)$$

Consistent with the prior correspondence in Eq. (2.288) used for the right-hand side, write for the left-hand side

$$L_r \phi_r - L_i \phi_i + i(L_r \phi_i + L_i \phi_r) \rightarrow \begin{bmatrix} L_r \phi_r - L_i \phi_i \\ L_r \phi_i + L_i \phi_r \end{bmatrix} = \begin{bmatrix} L_r & -L_i \\ L_i & L_r \end{bmatrix} \begin{bmatrix} \phi_r \\ \phi_i \end{bmatrix} \quad (2.308)$$

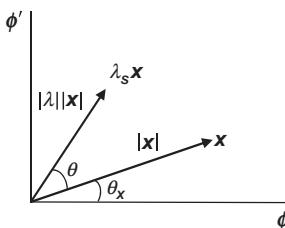


FIG. 2.29

Operation of spinor  $\lambda_s$ .

This correspondence thus replaces Eq. (2.288) in the  $N$ -dimensional complex space with

$$\begin{bmatrix} L_r & -L_i \\ L_i & L_r \end{bmatrix} \begin{bmatrix} \phi_r \\ \phi_i \end{bmatrix} = \begin{bmatrix} \lambda_r \phi_r - \lambda_i \phi_i \\ \lambda_r \phi_i + \lambda_i \phi_r \end{bmatrix} \quad (2.309)$$

in the  $2N$ -dimensional real space.

Using the previous results this relation may be rewritten as

$$\mathcal{L}\phi = \lambda_r \phi + \lambda_i \phi' \quad (2.310)$$

where

$$\mathcal{L} = \begin{bmatrix} L_r & -L_i \\ L_i & L_r \end{bmatrix} \quad (2.311)$$

Similarly

$$\mathcal{L}\phi' = \lambda_r \phi' - \lambda_i \phi \quad (2.312)$$

Now write

$$\lambda_r \phi + \lambda_i \phi' = (\lambda_r + \lambda_i \phi' \phi) \phi = \lambda_s \phi \quad (2.313)$$

$$\lambda_r \phi' - \lambda_i \phi = (\lambda_r + \lambda_i \phi' \phi) \phi' = \lambda_s \phi' \quad (2.314)$$

where  $\lambda_s = |\lambda| e^{i\phi\phi\theta}$  is the spinor defined before. Hence the eigenvectors  $\phi$  and  $\phi'$  share the same *eigenspinor*  $\lambda_s$ :

$$\mathcal{L}\phi = \lambda_s \phi \quad (2.315)$$

$$\mathcal{L}\phi' = \lambda_s \phi' \quad (2.316)$$

Therefore application of  $\mathcal{L}$  to any vector  $x$  in the plane shared by  $\phi$  and  $\phi'$  yields

$$\begin{aligned} \mathcal{L}x &= \mathcal{L}((x \cdot \phi)\phi + (x \cdot \phi')\phi') \\ &= (x \cdot \phi)\lambda_s \phi + (x \cdot \phi')\lambda_s \phi' \\ &= \lambda_s x \end{aligned} \quad (2.317)$$

As discussed before, this operation rotates and dilates  $x$  in the plane. This result shows that any vector in the plane shared by  $\phi$  and  $\phi'$  is an eigenvector of the linear operator  $\mathcal{L}$  with corresponding eigenspinor  $\lambda_s$ .

These considerations extend in a straightforward manner to an  $N \times N$  complex matrix  $L$  with  $N$  complex eigenvalues  $\lambda_{k,complex}, k = 1, \dots, N$  and  $N$  orthogonal complex eigenvectors  $\phi_{k,complex}, k = 1, \dots, N$ . To apply  $L$  to  $x_{complex}$  one first forms the corresponding real  $2N$ -dimensional problem as discussed before. One then projects the corresponding vector  $x$  onto  $N$  orthogonal *eigenbivectors* formed from the complex eigenvectors:

$$i_k = \phi'_k \wedge \phi_k = \phi'_k \phi_k \quad (2.318)$$

The corresponding eigenspinors  $\lambda_{s,k}, k = 1, \dots, N$  then rotate and dilate the resulting projected vectors in their respective planes. This operation may be written as

$$Lx_{complex} \rightarrow \mathcal{L}x = \sum_{k=1}^N \lambda_{s,k} [(x \cdot i_k) i_k^\dagger] \quad (2.319)$$

A geometric interpretation of transforming a complex vector by a complex matrix is therefore as follows:

1. Observe first that  $(\mathbf{x} \cdot i_k) i_k^\dagger$  is the orthogonal projection of the vector  $\mathbf{x}$  onto the two-dimensional subspace (plane) represented by the bivector  $i_k$ , which corresponds the complex eigenvector  $\boldsymbol{\phi}_{k,\text{complex}}$  of the complex matrix  $L$ . Thus complex eigenvectors correspond to eigenbivectors (i.e., planes).
2. Left-multiplication by the spinor  $\lambda_{s,k}$ , which is determined by the complex eigenvalue associated with the complex eigenvector  $\boldsymbol{\phi}_{k,\text{complex}}$ , then rotates and dilates this projected vector within the plane of the bivector  $i_k$ . Thus complex eigenvalues correspond to eigenspinors (i.e., rotations and dilations).

From this we see that a complex matrix with complex eigenvalues and complex eigenvectors corresponds to projecting a real vector onto orthogonal planes and then rotating and dilating the projected vectors. This geometric interpretation is analogous to the corresponding interpretation for an  $N \times N$  real matrix  $M$  with real eigenvalues  $\mu_k, k=1, \dots, N$  and real eigenvectors  $\mathbf{e}_k, k=1, \dots, N$ :

$$M\mathbf{z} = \sum_{k=1}^N \mu_k [(\mathbf{z} \cdot \mathbf{e}_k) \mathbf{e}_k] \quad (2.320)$$

In the real case,  $(\mathbf{z} \cdot \mathbf{e}_k) \mathbf{e}_k$  is orthogonal projection of the vector  $\mathbf{z}$  onto the one-dimensional subspace (line) represented by the eigenvector  $\mathbf{e}_k$ , and left multiplication by  $\mu_k$  “rotates” and dilates this projected vector where a “rotation” only occurs if multiplication by  $\mu_k$  results in a sign change.

This approach to understanding the operation of a complex matrix may be related to the traditional approach by first writing

$$\begin{aligned} \mathbf{x}_k &= (\mathbf{x} \cdot i_k) i_k^\dagger = (\mathbf{x} \cdot \boldsymbol{\phi}_k) \boldsymbol{\phi}_k + (\mathbf{x} \cdot \boldsymbol{\phi}'_k) \boldsymbol{\phi}'_k = ((\mathbf{x} \cdot \boldsymbol{\phi}_k) + (\mathbf{x} \cdot \boldsymbol{\phi}'_k) i_k) \boldsymbol{\phi}_k \\ &= |\mathbf{x}_k| e^{i_k \theta_{x,k}} \boldsymbol{\phi}_k \end{aligned} \quad (2.321)$$

where

$$|\mathbf{x}_k| = \sqrt{(\mathbf{x} \cdot \boldsymbol{\phi}_k)^2 + (\mathbf{x} \cdot \boldsymbol{\phi}'_k)^2} \quad (2.322)$$

$$\theta_{x,k} = \tan^{-1} \left( \frac{\mathbf{x} \cdot \boldsymbol{\phi}'_k}{\mathbf{x} \cdot \boldsymbol{\phi}_k} \right) \quad (2.323)$$

Also write

$$\lambda_{s,k} = |\lambda_k| e^{i_k \theta_k} \quad (2.324)$$

Hence

$$\begin{aligned} \lambda_{s,k} \mathbf{x}_k &= |\lambda_k| |\mathbf{x}_k| e^{i_k (\theta_k + \theta_{x,k})} \boldsymbol{\phi}_k \\ &= |\lambda_k| |\mathbf{x}_k| [\cos(\theta_k + \theta_{x,k}) \boldsymbol{\phi}_k + \sin(\theta_k + \theta_{x,k}) \boldsymbol{\phi}'_k] \end{aligned} \quad (2.325)$$

This relation yields

$$\mathcal{L}\mathbf{x} = \sum_{k=1}^N |\lambda_k| |\mathbf{x}_k| [\cos(\theta_k + \theta_{x,k}) \boldsymbol{\phi}_k + \sin(\theta_k + \theta_{x,k}) \boldsymbol{\phi}'_k] \quad (2.326)$$

This may be rewritten as

$$\mathcal{L}\mathbf{x} = \sum_{k=1}^N \left\{ \operatorname{Re} [\lambda_{k,\text{complex}} x_{k,\text{complex}}] \boldsymbol{\phi}_k + \operatorname{Im} [\lambda_{k,\text{complex}} x_{k,\text{complex}}] \boldsymbol{\phi}'_k \right\} \quad (2.327)$$

where

$$|\lambda_k| |\mathbf{x}_k| \cos(\theta_k + \theta_{x,k}) = \operatorname{Re} [\lambda_{k,\text{complex}} x_{k,\text{complex}}] \quad (2.328)$$

$$|\lambda_k| |\mathbf{x}_k| \sin(\theta_k + \theta_{x,k}) = \operatorname{Im} [\lambda_{k,\text{complex}} x_{k,\text{complex}}] \quad (2.329)$$

In these expressions  $\lambda_{k,\text{complex}}$  is the complex eigenvalue with corresponding complex eigenvector  $\boldsymbol{\phi}_{k,\text{complex}}$  and  $x_{k,\text{complex}} = (\mathbf{x}_{\text{complex}} \cdot \boldsymbol{\phi}_{k,\text{complex}})$  is a complex number. The correspondence in Eq. (2.288) now leads to

$$\lambda_{k,\text{complex}} x_{k,\text{complex}} \boldsymbol{\phi}_{k,\text{complex}} \rightarrow \operatorname{Re} [\lambda_{k,\text{complex}} x_{k,\text{complex}}] \boldsymbol{\phi}_k + \operatorname{Im} [\lambda_{k,\text{complex}} x_{k,\text{complex}}] \boldsymbol{\phi}'_k \quad (2.330)$$

Hence

$$\begin{aligned} \mathcal{L}\mathbf{x} &\rightarrow \mathcal{L}\mathbf{x}_{\text{complex}} \\ &= \sum_{k=1}^N \lambda_{k,\text{complex}} x_{k,\text{complex}} \boldsymbol{\phi}_{k,\text{complex}} \\ &= \sum_{k=1}^N \lambda_{k,\text{complex}} (\mathbf{x}_{\text{complex}} \cdot \boldsymbol{\phi}_{k,\text{complex}}) \boldsymbol{\phi}_{k,\text{complex}} \end{aligned} \quad (2.331)$$

This is the traditional formulation of the eigen-expansion using complex numbers and vectors. However, the traditional approach does not yield the geometric interpretation given earlier.

### 2.2.5.1 Geometry of the Matrix Inverse

A significant aspect of radar signal processing encompasses computing the inverse of the covariance matrix of the observed data. See Ref. [50] for an example of an important application where this occurs. Thus it is also of some interest to examine the geometric effect of applying the linear transformation defined by this inverse matrix to vectors in a  $2N$ -dimensional space.

To this end we begin by considering an  $N$ -dimensional vector  $\mathbf{z}$  and examine the operation of the linear operator

$$f(\mathbf{a}) = \mathbf{z}\mathbf{z}^T \mathbf{a} \quad (2.332)$$

where  $\mathbf{a}$  is an arbitrary vector. Let us write  $\mathbf{z}^T \mathbf{a} = \mathbf{z} \cdot \mathbf{a}$  and let  $\{\boldsymbol{\varphi}_i, i=1, \dots, N\}$  be a basis for the  $N$ -dimensional space. Note that we are not assuming that the basis is orthogonal. Let us now expand  $\mathbf{z}$  in this basis:

$$\mathbf{z} = \sum_{i=1}^N z_{i,\varphi} \boldsymbol{\varphi}_i \quad (2.333)$$

with

$$z_{i,\varphi} = \mathbf{z} \cdot \boldsymbol{\varphi}_i. \quad (2.334)$$

It is important to note that  $z_{i,\varphi}$  depends on the basis that is chosen. With this expansion, Eq. (2.332) becomes

$$\begin{aligned} f(\mathbf{a}) &= \left( \sum_{i=1}^N z_{i,\varphi} \boldsymbol{\varphi}_i \right) \left( \sum_{j=1}^N z_{j,\varphi} (\boldsymbol{\varphi}_j \cdot \mathbf{a}) \right) \\ &= \sum_{i=1}^N z_{i,\varphi}^2 (\boldsymbol{\varphi}_i \cdot \mathbf{a}) \boldsymbol{\varphi}_i + \sum_{i=1}^N \sum_{\substack{j=1 \\ j \neq i}}^N z_{i,\varphi} z_{j,\varphi} (\boldsymbol{\varphi}_j \cdot \mathbf{a}) \boldsymbol{\varphi}_i. \end{aligned} \quad (2.335)$$

Now observe that a covariance matrix for a zero mean vector  $\mathbf{z}$  is given by

$$R = E[\mathbf{z}\mathbf{z}^T]. \quad (2.336)$$

Thus treating  $R$  as a linear operator leads to

$$Ra = \sum_{i=1}^N \lambda_{i,\varphi} (\boldsymbol{\varphi}_i \cdot \mathbf{a}) \boldsymbol{\varphi}_i + \sum_{i=1}^N \sum_{\substack{j=1 \\ j \neq i}}^N C_{ij,\varphi} (\boldsymbol{\varphi}_j \cdot \mathbf{a}) \boldsymbol{\varphi}_i \quad (2.337)$$

where

$$\lambda_{i,\varphi} = E[z_{i,\varphi}^2] \quad (2.338)$$

$$C_{ij,\varphi} = E[z_{i,\varphi} z_{j,\varphi}]. \quad (2.339)$$

Again, it is important to note that these quantities depend on the chosen basis. As a first example, let us choose the *observed data basis*, i.e.

$$\boldsymbol{\varphi}_i = \boldsymbol{\varphi}_{data,i} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad (2.340)$$

where the 1 appears in the  $i$ th position. This choice yields

$$z_{i,\varphi} = \mathbf{z} \cdot \boldsymbol{\varphi}_i = z_i \quad (2.341)$$

where

$$\mathbf{z} = \begin{bmatrix} z_1 \\ \vdots \\ z_N \end{bmatrix} \quad (2.342)$$

gives the vector  $\mathbf{z}$  in terms of the components of the collected data. If we assume zero-mean stationary data, then in this basis we find

$$\lambda_{i,\varphi} = E[z_i^2] = \sigma^2, \quad i = 1, \dots, N \quad (2.343)$$

and

$$C_{ij,\varphi} = E[z_i z_j] = \sigma^2 \rho_{ij} \quad (2.344)$$

where  $\sigma^2$  denotes the variance of the stationary process and  $\rho_{ij}$  gives the correlation coefficient between  $z_i$  and  $z_j$ .

The representation of the covariance matrix as a linear operator then becomes in this basis

$$\begin{aligned} R\mathbf{a} &= \sigma^2 \sum_{i=1}^N (\boldsymbol{\varphi}_{data,i} \cdot \mathbf{a}) \boldsymbol{\varphi}_{data,i} + \sigma^2 \sum_{i=1}^N \sum_{\substack{j=1 \\ j \neq i}}^N \rho_{ij} (\boldsymbol{\varphi}_{data,j} \cdot \mathbf{a}) \boldsymbol{\varphi}_{data,i} \\ &= \sigma^2 \left( \mathbf{a} + \sum_{i=1}^N \sum_{\substack{j=1 \\ j \neq i}}^N \rho_{ij} (\boldsymbol{\varphi}_{data,j} \cdot \mathbf{a}) \boldsymbol{\varphi}_{data,i} \right). \end{aligned} \quad (2.345)$$

Observe that when  $\rho_{ij}=0$  we obtain  $R\mathbf{a}=\sigma^2\mathbf{a}$ , i.e.,  $R=\sigma^2 I$ , which is what we expect. Writing out the expression in Eq. (2.345) shows that in general it is equivalent to the ordinary matrix multiplication of  $R$  times  $\mathbf{a}$  where

$$R = \sigma^2 \begin{bmatrix} \rho_{11} & \dots & \rho_{1N} \\ \vdots & \ddots & \vdots \\ \rho_{N1} & \dots & \rho_{NN} \end{bmatrix} \quad (2.346)$$

and  $\rho_{ii}=1$ ,  $i=1, \dots, N$ .

Instead of the observed data basis, consider now a different basis, designated  $\boldsymbol{\varphi}_i = \boldsymbol{\varphi}_{e,i}$ , with the following properties:

$$\lambda_{i,\varphi} = E[z_{i,e}^2] = E[(\mathbf{z} \cdot \boldsymbol{\varphi}_{e,i})^2] = \lambda_i, \quad i = 1, \dots, N \quad (2.347)$$

$$C_{ij,\varphi} = E[z_{i,e} z_{j,e}] = E[(\mathbf{z} \cdot \boldsymbol{\varphi}_{e,i})(\mathbf{z} \cdot \boldsymbol{\varphi}_{e,j})] = 0. \quad (2.348)$$

With this basis, Eq. (2.337) becomes

$$R\mathbf{a} = \sum_{i=1}^N \lambda_i (\boldsymbol{\varphi}_{e,i} \cdot \mathbf{a}) \boldsymbol{\varphi}_{e,i}. \quad (2.349)$$

This is the eigen-expansion of the linear operator  $R$ , which is easy to show by setting  $R\boldsymbol{\varphi}_{e,i} = \lambda_i \boldsymbol{\varphi}_{e,i}$  and expanding  $\mathbf{a}$  in this basis. From this development, we see:

1. The eigenvalue  $\lambda_i$  represents the variance of the process in the direction defined by the eigenvector  $\boldsymbol{\varphi}_{e,i}$ , and
2. The eigenvectors designate directions such that the process has no correlation between the different eigenvector directions.

In this basis, we recover the case  $R=\sigma^2 I$  by setting  $\lambda_i=\sigma^2$  for all  $i=1, \dots, N$ . Thus the effect of correlation is seen to change the variance of the process in the different directions defined by the eigenvectors. The eigen-expansion of the linear

operator represented by the covariance matrix  $R$  is particularly convenient in that the inverse operator is easily defined by simply inverting the eigenvalues:

$$R^{-1}\mathbf{a} = \sum_{i=1}^N \frac{1}{\lambda_i} (\boldsymbol{\varphi}_{e,i} \cdot \mathbf{a}) \boldsymbol{\varphi}_{e,i}. \quad (2.350)$$

To apply these ideas to in-phase and quadrature data, assume now that the covariance matrix of the complex noise data is a complex Hermitian, positive-semidefinite  $N \times N$  matrix. Thus it has  $N$  complex eigenvectors with real eigenvalues. Let us write for the  $i$ th eigenvalue  $\lambda_i$

$$R\boldsymbol{\varphi}_{i,\text{complex}} = \lambda_i \boldsymbol{\varphi}_{i,\text{complex}} \quad (2.351)$$

where

$$\boldsymbol{\varphi}_{i,\text{complex}} = \boldsymbol{\varphi}_{i,\text{real}} + j\boldsymbol{\varphi}_{i,\text{imag}} \quad (2.352)$$

is the associated complex eigenvector and the eigenvectors satisfy:

$$\boldsymbol{\varphi}_{i,\text{complex}}^* \cdot \boldsymbol{\varphi}_{k,\text{complex}} = 0, \quad i \neq k. \quad (2.353)$$

As before, from each complex eigenvector form two new vectors in the  $2N$ -dimensional space:

$$\boldsymbol{\varphi}_i = \begin{bmatrix} \boldsymbol{\varphi}_{i,\text{real}} \\ \boldsymbol{\varphi}_{i,\text{imag}} \end{bmatrix}, \boldsymbol{\varphi}'_i = \begin{bmatrix} -\boldsymbol{\varphi}_{i,\text{imag}} \\ \boldsymbol{\varphi}_{i,\text{real}} \end{bmatrix}, \quad i = 1, \dots, N \quad (2.354)$$

It follows from this formulation that  $\boldsymbol{\varphi}_i, \boldsymbol{\varphi}'_i, i = 1, \dots, N$  form an orthonormal basis in the  $2N$ -dimensional space. Now form the following  $2N \times 2N$  covariance matrix

$$\mathcal{R} = \begin{bmatrix} R_{\text{real}} & -R_{\text{imag}} \\ R_{\text{imag}} & R_{\text{real}} \end{bmatrix} \quad (2.355)$$

where  $R = R_{\text{real}} + jR_{\text{imag}}$  is the  $N \times N$  covariance matrix for the complex noise samples. Finally, denote a complex vector  $\mathbf{z}$  as

$$\mathbf{z} = \mathbb{I} + i\mathbb{Q} \quad (2.356)$$

$$\mathbf{d} = \begin{bmatrix} \mathbb{I} \\ \mathbb{Q} \end{bmatrix} \quad (2.357)$$

where  $\mathbb{I}, \mathbb{Q}$  are real  $N$ -dimensional vectors corresponding to the in-phase and quadrature part of  $\mathbf{z}$  (in Eq. 2.356,  $i = \sqrt{-1}$ .)

This notation leads to

$$\mathcal{R}\mathbf{d} = \begin{bmatrix} R_{\text{real}} & -R_{\text{imag}} \\ R_{\text{imag}} & R_{\text{real}} \end{bmatrix} \begin{bmatrix} \mathbb{I} \\ \mathbb{Q} \end{bmatrix} = \begin{bmatrix} R_{\text{real}}\mathbb{I} - R_{\text{imag}}\mathbb{Q} \\ R_{\text{imag}}\mathbb{I} + R_{\text{real}}\mathbb{Q} \end{bmatrix} \rightarrow R\mathbf{z} \quad (2.358)$$

where this latter notation indicates that the vector  $\mathcal{R}\mathbf{d}$  corresponds to the vector  $R\mathbf{z}$  in the same way that  $\mathbf{d}$  corresponds to  $\mathbf{z}$ . From this formulation it follows that

$$\mathcal{R}\boldsymbol{\varphi}_i = \lambda_i \boldsymbol{\varphi}_i \quad (2.359)$$

$$\mathcal{R}\boldsymbol{\varphi}'_i = \lambda_i \boldsymbol{\varphi}'_i. \quad (2.360)$$

Thus the  $2N \times 2N$  covariance matrix  $\mathcal{R}$  has  $2N$  real eigenvectors  $\boldsymbol{\varphi}_i$  and  $\boldsymbol{\varphi}'_i, i = 1, \dots, N$  with corresponding real eigenvalues  $\lambda_i, i = 1, \dots, N$  where each eigenvalue  $\lambda_i$  is associated with two eigenvectors.

Now examine the bivectors formed from the pairs of eigenvectors sharing the same eigenvalue

$$i_k = \boldsymbol{\varphi}_k \wedge \boldsymbol{\varphi}'_k \quad (2.361)$$

As shown before, with this result we may expand  $\mathcal{R}\mathbf{d}$  in an eigenbivector expansion:

$$\mathcal{R}\mathbf{d} = \sum_{k=1}^N \lambda_k \left[ (\mathbf{d} \cdot i_k) i_k^\dagger \right] \quad (2.362)$$

It follows that

$$R^{-1}\mathbf{z} \rightarrow \mathcal{R}^{-1}\mathbf{d} = \sum_{k=1}^N \frac{1}{\lambda_k} \left[ (\mathbf{d} \cdot i_k) i_k^\dagger \right] \quad (2.363)$$

Because the bivectors  $i_k, k = 1, \dots, N$  are mutually orthogonal, it follows that the vector obtained by operating on a data vector  $\mathbf{z}$  with the matrix inverse  $R^{-1}$  may be expressed as an expansion in weighted, mutually orthogonal real vectors obtained by orthogonally projecting the data onto the set of  $N$  eigenbivectors of  $\mathcal{R}$  that span the  $2N$ -dimensional space and then weighting them by the inverse of the eigenvalues associated with those eigenbivectors. Moreover, we see from Eq. (2.363) that prior to projecting the data onto the steering bivector the effect of this linear transformation is to deemphasize the part of the data vector  $\mathbf{d}$  that lies in the subspaces defined by the bivectors associated with the largest eigenvalues of  $R$ .

From this result, we see that a quantity such as

$$\mathbf{s}^* \cdot R^{-1}\mathbf{z} \quad (2.364)$$

has the same geometric interpretation as  $\mathbf{s}^* \cdot \mathbf{z}$  (to be discussed in the next section) except that the linear operation  $R^{-1}\mathbf{z}$  acts to deemphasize the components of the data  $\mathbf{z}$  in the subspaces corresponding to the largest eigenvalues of  $R$ .

## 2.3 SELECTED APPLICATIONS TO RADAR SIGNAL PROCESSING

### 2.3.1 HERMITIAN INNER PRODUCT

Consider the Hermitian inner product

$$\mathbf{s}^* \cdot \mathbf{z} \quad (2.365)$$

where  $\mathbf{s}$  and  $\mathbf{z}$  are the following  $N$ -dimensional complex vectors:

$$\mathbf{z} = \begin{bmatrix} \mathcal{J}_0 + iQ_0 \\ \vdots \\ \mathcal{J}_{N-1} + iQ_{N-1} \end{bmatrix} = \mathbb{I} + i\mathbb{Q} \quad (2.366)$$

$$\mathbf{s} = \begin{bmatrix} \mathcal{J}_{p,0} + i\mathcal{Q}_{p,0} \\ \vdots \\ \mathcal{J}_{p,N-1} + i\mathcal{Q}_{p,N-1} \end{bmatrix} = \mathbb{I}_p + i\mathbb{Q}_p \quad (2.367)$$

In these expressions  $\mathbb{I}, \mathbb{Q}, \mathbb{I}_p, \mathbb{Q}_p$  are  $N$ -dimensional real vectors. Such expressions are ubiquitous in radar signal processing and appear in such contexts as matched filtering, angle estimation in an array, and space–time adaptive processing. The complex vector  $\mathbf{s}$  is often called a *steering vector*; it can be quite general and can represent such quantities as a Doppler vector whose components capture the pulse-to-pulse Doppler rotation of a frequency-shifted signal or an array steering vector whose components define a beam or null steering direction utilized in array processing. A straightforward computation yields

$$\begin{aligned} \mathbf{s}^* \cdot \mathbf{z} &= \operatorname{Re}[\mathbf{s}^* \cdot \mathbf{z}] + i \operatorname{Im}[\mathbf{s}^* \cdot \mathbf{z}] \\ &= \mathbb{I}_p \cdot \mathbb{I} + \mathbb{Q}_p \cdot \mathbb{Q} - i(\mathbb{Q}_p \cdot \mathbb{I} - \mathbb{I}_p \cdot \mathbb{Q}) \end{aligned} \quad (2.368)$$

Associate with  $\mathbf{s}$  and  $\mathbf{z}$  the following  $2N$ -dimensional real vectors:

$$\mathbf{p} = \begin{bmatrix} \mathbb{I}_p \\ \mathbb{Q}_p \end{bmatrix} \quad (2.369)$$

$$\mathbf{p}' = \begin{bmatrix} -\mathbb{Q}_p \\ \mathbb{I}_p \end{bmatrix} \quad (2.370)$$

$$\mathbf{d} = \begin{bmatrix} \mathbb{I} \\ \mathbb{Q} \end{bmatrix} \quad (2.371)$$

We will assume a normalization such that  $|\mathbf{s}| = |\mathbf{p}| = |\mathbf{p}'| = 1$ . We also have

$$\mathbf{p} \cdot \mathbf{p}' = 0. \quad (2.372)$$

Thus  $\mathbf{p}$  and  $\mathbf{p}'$  are orthonormal vectors that form a two-dimensional subspace (i.e., a plane) in the  $2N$ -dimensional space. In this notation we may write

$$\mathbf{s}^* \cdot \mathbf{z} = \mathbf{p} \cdot \mathbf{d} - i\mathbf{p}' \cdot \mathbf{d}. \quad (2.373)$$

Now define the unit bivector  $i_p$  in the  $\mathbf{p}, \mathbf{p}'$ -plane:

$$i_p = \mathbf{p}' \wedge \mathbf{p} = \mathbf{p}' \mathbf{p}$$

From the relation in Eq. (2.373),  $\mathbf{s}^* \cdot \mathbf{z}$  is evidently related to the projection of the vector  $\mathbf{d}$  into the subspace (i.e., plane) defined by  $\mathbf{p}$  and  $\mathbf{p}'$ :

$$\mathbf{d}_p = (\mathbf{d} \cdot i_p) i_p^\dagger = (\mathbf{p} \cdot \mathbf{d}) \mathbf{p} + (\mathbf{p}' \cdot \mathbf{d}) \mathbf{p}' \quad (2.374)$$

To relate  $\mathbf{d}_p$  to  $\mathbf{s}^* \cdot \mathbf{z}$ , multiply  $\mathbf{d}_p$  from the left by  $\mathbf{p}$ :

$$\begin{aligned} \mathbf{p} \mathbf{d}_p &= (\mathbf{p} \cdot \mathbf{d}) \mathbf{p} \mathbf{p} + (\mathbf{p}' \cdot \mathbf{d}) \mathbf{p} \mathbf{p}' \\ &= \mathbf{p} \cdot \mathbf{d} - i_p \mathbf{p}' \cdot \mathbf{d} \end{aligned} \quad (2.375)$$

Comparison of Eqs. (2.373) and (2.375) show that by interpreting the unit imaginary  $i$  as the bivector  $i_p$ , the Hermitian inner product  $\mathbf{s}^* \cdot \mathbf{z}$  can be interpreted as the

geometric product of two  $2N$ -dimensional real vectors that represent the steering signal and the data:

$$\mathbf{s}^* \cdot \mathbf{z} = \mathbf{p} \mathbf{d}_p \quad (2.376)$$

Thus  $\mathbf{s}^* \cdot \mathbf{z}$  is a measure of the relative magnitudes of the real vectors  $\mathbf{p}$  and  $\mathbf{d}_p$  and the angle between them in the  $\mathbf{p}, \mathbf{p}'$  plane. Since  $|\mathbf{p}|=1$  and since  $\mathbf{d}_p$  is the orthogonal projection of  $\mathbf{d}$  onto the  $\mathbf{p}, \mathbf{p}'$  plane, it follows that  $\mathbf{s}^* \cdot \mathbf{z}$  is a measure of the magnitude of  $\mathbf{d}$  and the angle between  $\mathbf{p}$  and  $\mathbf{d}_p$  in the  $\mathbf{p}, \mathbf{p}'$  plane.

Fig. 2.30 shows the relationships between the various vectors in the  $2N$ -dimensional space.

In this picture, the vector  $\mathbf{d}_p$  is in the  $\mathbf{p}, \mathbf{p}'$  plane and is obtained by orthogonally projecting  $\mathbf{d}$  through an angle  $\psi_p$  onto the  $\mathbf{p}, \mathbf{p}'$  plane, i.e., onto  $i_p$ . Thus we have

$$\mathbf{p} \cdot \mathbf{d} = \mathbf{p} \cdot \mathbf{d}_p = |\mathbf{p}| |\mathbf{d}_p| \cos \theta_p \quad (2.377)$$

$$\mathbf{p}' \cdot \mathbf{d} = \mathbf{p}' \cdot \mathbf{d}_p = |\mathbf{p}'| |\mathbf{d}_p| \sin \theta_p \quad (2.378)$$

It is evident that

$$\begin{aligned} \mathbf{s}^* \cdot \mathbf{z} &= |\mathbf{p}| |\mathbf{d}| \cos \psi_p \{ \cos \theta_p - i_p \sin \theta_p \} \\ &= |\mathbf{d}| \cos \psi_p e^{-i_p \theta_p} \end{aligned} \quad (2.379)$$

Here we have used  $|\mathbf{p}|=1$  and recognized that  $|\mathbf{d}_p|=|\mathbf{d}| \cos \psi_p$ . Therefore Eq. (2.379) shows that the Hermitian inner product of two complex vectors may also be interpreted as a spinor in geometric algebra. The magnitude of the vector  $\mathbf{d}_p$ —and thus the magnitude of the spinor  $\mathbf{s}^* \cdot \mathbf{z}$ —is independent of the choice of the specific orthonormal basis  $\mathbf{p}$  and  $\mathbf{p}'$  in the plane (they may be rotated arbitrarily in this plane). The angle  $\theta_p$ , on the other hand, is a function of how these axes are chosen.

In computing  $\mathbf{s}^* \cdot \mathbf{z}$ , the first step is to obtain  $\mathbf{d}_p$  by projecting  $\mathbf{d}$  onto  $i_p$ . A very important example is the case where

$$\begin{aligned} \mathbf{z} &= Ae^{i\phi}\mathbf{s} + \mathbf{n}_{\text{complex}} \\ &= Ae^{i\phi}(\mathbb{I}_p + i\mathbb{Q}_p) + \mathbb{I}_{\text{noise}} + i\mathbb{Q}_{\text{noise}} \\ &= A(\cos \phi \mathbb{I}_p - \sin \phi \mathbb{Q}_p) + iA(\cos \phi \mathbb{Q}_p + \sin \phi \mathbb{I}_p) + \mathbb{I}_{\text{noise}} + i\mathbb{Q}_{\text{noise}} \end{aligned} \quad (2.380)$$

Assume for now that the phase offset  $\phi$  is known. Then we may define a new steering vector

$$\mathbf{s}_\phi = e^{i\phi}\mathbf{s} = (\cos \phi \mathbb{I}_p - \sin \phi \mathbb{Q}_p) + i(\cos \phi \mathbb{Q}_p + \sin \phi \mathbb{I}_p) \quad (2.381)$$

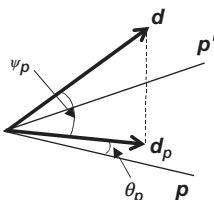


FIG. 2.30

Projection of data vector onto plane defined by steering vector.

Thus

$$z = As_\phi + \mathbf{n}_{\text{complex}} \quad (2.382)$$

The data vector  $\mathbf{d}$  corresponding to  $z$  becomes

$$\mathbf{d} = A\mathbf{p}_\phi + \mathbf{n} \quad (2.383)$$

where  $\mathbf{n}$  represents the  $2N$ -dimensional real noise vector corresponding to the  $N$ -dimensional complex vector  $\mathbf{n}_{\text{complex}}$ . The corresponding data vectors  $\mathbf{p}_\phi, \mathbf{p}'_\phi$  in the  $2N$ -dimensional space are given by

$$\begin{aligned} \mathbf{p}_\phi &= \begin{bmatrix} \cos\phi \mathbb{I}_p - \sin\phi \mathbb{Q}_p \\ \cos\phi \mathbb{Q}_p + \sin\phi \mathbb{I}_p \end{bmatrix} = \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix} \begin{bmatrix} \mathbb{I}_p \\ \mathbb{Q}_p \end{bmatrix} \\ &= \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix} \mathbf{p} \end{aligned} \quad (2.384)$$

$$\begin{aligned} \mathbf{p}'_\phi &= \begin{bmatrix} -\sin\phi \mathbb{I}_p - \cos\phi \mathbb{Q}_p \\ \cos\phi \mathbb{I}_p - \sin\phi \mathbb{Q}_p \end{bmatrix} = \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix} \begin{bmatrix} -\mathbb{Q}_p \\ \mathbb{I}_p \end{bmatrix} \\ &= \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix} \mathbf{p}' \end{aligned} \quad (2.385)$$

These are simply rotations of  $\mathbf{p}$  and  $\mathbf{p}'$  in the  $\mathbf{p}, \mathbf{p}'$  plane through the angle  $\phi$  as shown in Fig. 2.31.

The vectors  $\mathbf{p}_\phi, \mathbf{p}'_\phi$  also form an orthonormal basis in the plane. As a result we have

$$i_p = \mathbf{p}' \mathbf{p} = \mathbf{p}'_\phi \mathbf{p}_\phi \quad (2.386)$$

As mentioned before, rotating  $\mathbf{p}$  and  $\mathbf{p}'$  to obtain  $\mathbf{p}_\phi$  and  $\mathbf{p}'_\phi$  does not affect the unit bivector  $i_p$ . Now examine the orthogonally projected vector  $\mathbf{d}_p$  in this new coordinate system:

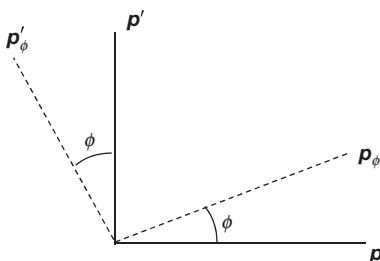


FIG. 2.31

Rotation of steering vector by a known phase offset.

$$\begin{aligned}
\mathbf{d}_p &= (\mathbf{d} \cdot i_p) i_p^\dagger = (\mathbf{p}_\phi \cdot \mathbf{d}) \mathbf{p}_\phi + (\mathbf{p}'_\phi \cdot \mathbf{d}) \mathbf{p}'_\phi \\
&= (\mathbf{p}_\phi \cdot (A \mathbf{p}_\phi + \mathbf{n})) \mathbf{p}_\phi + (\mathbf{p}'_\phi \cdot (A \mathbf{p}_\phi + \mathbf{n})) \mathbf{p}'_\phi \\
&= A \mathbf{p}_\phi + (\mathbf{p}_\phi \cdot \mathbf{n}) \mathbf{p}_\phi + (\mathbf{p}'_\phi \cdot \mathbf{n}) \mathbf{p}'_\phi
\end{aligned} \tag{2.387}$$

In the absence of noise, the projected vector  $\mathbf{d}_p$  would be given by  $A \mathbf{p}_\phi$ , and the magnitude  $A$  of this projected vector would be given by

$$A = \operatorname{Re}[\mathbf{s}_\phi^* \cdot \mathbf{z}] = \mathbf{p}_\phi \cdot \mathbf{d} \tag{2.388}$$

In this case  $\mathbf{s}_\phi^* \cdot \mathbf{z}$  may be interpreted as the projection of the data vector  $\mathbf{d}$  onto a one-dimensional subspace  $\mathbf{p}_\phi$ . The rotated steering vector  $\mathbf{p}_\phi$  has a meaningful interpretation as the direction of the signal represented by  $\mathbf{d}$  in the absence of noise.

However, in this same example, if the initial phase offset is unknown, then the coordinate system  $\mathbf{p}_\phi$  and  $\mathbf{p}'_\phi$  cannot be determined and the direction of the projected vector  $\mathbf{d}_p$  in the plane relative to  $\mathbf{p}$  and  $\mathbf{p}'$  is unknown. In this case the Hermitian inner product  $\mathbf{s}^* \cdot \mathbf{z}$ , which in general has a magnitude as well as an angle measured relative to  $\mathbf{p}$ , is obtained by projecting the data vector  $\mathbf{d}$  into a two-dimensional subspace  $i_p$  rather than a one-dimensional subspace  $\mathbf{p}_\phi$ . Unlike in the case where the phase offset is known, when the phase offset is unknown, the vector  $\mathbf{p}$ , which differs from  $\mathbf{p}_\phi$ , may have no discernible relationship to the direction of the signal represented by  $\mathbf{d}$ . Also, examine

$$\begin{aligned}
\operatorname{Re}[\mathbf{s}^* \cdot \mathbf{z}] &= \mathbf{p} \cdot \mathbf{d} = A \mathbf{p} \cdot \mathbf{p}_\phi + \mathbf{p} \cdot \mathbf{n} \\
&= A \cos \phi + \mathbf{p} \cdot \mathbf{n}
\end{aligned} \tag{2.389}$$

In this case, even in the absence of noise, the quantity  $\operatorname{Re}[\mathbf{s}^* \cdot \mathbf{z}]$  bears no useful relationship to the amplitude of the signal since  $\cos \phi$  can take on any value in  $[-1, 1]$ . Instead examine

$$\begin{aligned}
|\mathbf{s}^* \cdot \mathbf{z}| &= |\mathbf{d}_p| = \sqrt{(\mathbf{p} \cdot \mathbf{d})^2 + (\mathbf{p}' \cdot \mathbf{d})^2} \\
&= \sqrt{(A \mathbf{p} \cdot \mathbf{p}_\phi + \mathbf{p} \cdot \mathbf{n})^2 + (A \mathbf{p}' \cdot \mathbf{p}_\phi + \mathbf{p}' \cdot \mathbf{n})^2} \\
&= \sqrt{(A \cos \phi + \mathbf{p} \cdot \mathbf{n})^2 + (A \sin \phi + \mathbf{p}' \cdot \mathbf{n})^2}
\end{aligned} \tag{2.390}$$

In the absence of noise, this relation becomes

$$|\mathbf{s}^* \cdot \mathbf{z}| = A \tag{2.391}$$

Thus when the phase offset is unknown, which is a common case, the amplitude of the signal in this example is more usefully determined by the magnitude of  $\mathbf{s}^* \cdot \mathbf{z}$ , not just its real part.

More generally, whether  $\phi$  is known or not, in this example (i.e., Eqs. 2.383 and 2.386) we have

$$\begin{aligned}
\mathbf{d}_p &= (\mathbf{d} \cdot i_p) i_p^\dagger = ((A \mathbf{p}_\phi + \mathbf{n}) \cdot i_p) i_p^\dagger \\
&= A (\mathbf{p}_\phi \cdot i_p) i_p^\dagger + (\mathbf{n} \cdot i_p) i_p^\dagger
\end{aligned} \tag{2.392}$$

Examine

$$\begin{aligned} (\mathbf{p}_\phi \cdot i_p) i_p^\dagger &= \left( (\mathbf{p}_\phi \cdot \mathbf{p}'_\phi) \mathbf{p}_\phi - (\mathbf{p}_\phi \cdot \mathbf{p}_\phi) \mathbf{p}'_\phi \right) \mathbf{p}_\phi \mathbf{p}'_\phi \\ &= -\mathbf{p}'_\phi \mathbf{p}_\phi \mathbf{p}'_\phi \\ &= \mathbf{p}_\phi \end{aligned} \quad (2.393)$$

This result occurs because  $\mathbf{p}_\phi$  is in the plane corresponding to  $i_p$  and therefore projecting it into this subspace simply returns the same vector. Also, denote the projection of the noise vector into this plane as

$$\mathbf{n}_p = (\mathbf{n} \cdot i_p) i_p^\dagger \quad (2.394)$$

Then we have for the projected vector  $\mathbf{d}_p$ :

$$\mathbf{d}_p = A\mathbf{p}_\phi + \mathbf{n}_p \quad (2.395)$$

Under the assumption of zero-mean white Gaussian noise we have

$$\langle \mathbf{d}_p \rangle = A\mathbf{p}_\phi \quad (2.396)$$

In accordance with the orthogonality principle [52], the vector  $\mathbf{d}_p$  is now seen to be the minimum mean-square estimate of the signal vector  $A\mathbf{p}_\phi = Ae^{i\phi}\mathbf{s}$ . Thus in this example the Hermitian inner product is related to the minimum mean-square estimate of the underlying signal.

From Eq. (2.376) we obtain

$$\mathbf{d}_p = \mathbf{p}(s^* \cdot z) \quad (2.397)$$

As discussed before  $s^* \cdot z$  is a spinor in the  $\mathbf{p}, \mathbf{p}'$ -plane. Therefore it follows from the discussion about rotations in Section 2.2.3 that this relation may also be written as

$$\mathbf{d}_p = (s^* \cdot z)^* \mathbf{p} \quad (2.398)$$

Geometrically, therefore, the Hermitian inner product becomes

$$s^* \cdot z = \mathbf{p} \mathbf{d}_p \quad (2.399)$$

$$(s^* \cdot z)^* = \mathbf{d}_p \mathbf{p} \quad (2.400)$$

The Hermitian inner product  $(s^* \cdot z)^*$  is a spinor that rotates the steering vector  $\mathbf{p}$  in the  $\mathbf{p}, \mathbf{p}'$ -plane to obtain an estimate of the signal given by  $\mathbf{d}_p$ .

To state this result in terms of the original complex vectors  $\mathbf{s}$  and  $\mathbf{z}$ , examine

$$\mathbf{p} \cdot \mathbf{d} = Re[s^* \cdot z] \quad (2.401)$$

$$\mathbf{p}' \cdot \mathbf{d} = -Im[s^* \cdot z] \quad (2.402)$$

Writing  $\mathbf{d}_p$  in terms of these relations, obtaining the corresponding complex vector, and treating the complex vector corresponding to  $\mathbf{d}_p$  as an estimate of  $Ae^{i\phi}\mathbf{s}$  yields

$$A\hat{e}^{i\phi} = (s^* \cdot z)^* = \mathbf{d}_p \mathbf{p} \quad (2.403)$$

This relation leads to the interpretation of the Hermitian inner product  $\mathbf{s}^* \cdot \mathbf{z}$  in this example as an estimate of the complex amplitude  $Ae^{-i\phi}$ . This interpretation is consistent with the spinor interpretation in Eq. (2.379) with

$$\hat{\phi} = \theta_p \quad (2.404)$$

$$\hat{A} = |\mathbf{d}_p| = |\mathbf{s}^* \cdot \mathbf{z}| \quad (2.405)$$

### 2.3.2 THE GEOMETRY OF SIGNAL DETECTION

#### 2.3.2.1 Multivariate Gaussian PDF and a Simple Detection Problem

Consider a univariate Gaussian random variable  $X$  with zero mean and unit variance. Its probability density function (PDF) is well known and is given by

$$p_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \quad (2.406)$$

Now consider  $N$  independent and identically distributed random variables  $X_i$ ,  $i = 1, \dots, N$ . The multivariate PDF is also well known and is given by

$$p_{X_1, \dots, X_N}(x_1, \dots, x_N) = \frac{1}{(\sqrt{2\pi})^N} e^{-\frac{x_1^2 + \dots + x_N^2}{2}} \quad (2.407)$$

Let us organize these random variables into a random vector

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_N \end{bmatrix} \quad (2.408)$$

Then we may write

$$p_X(\mathbf{x}) = \frac{1}{(\sqrt{2\pi})^N} e^{-\frac{\mathbf{x}^2}{2}} \quad (2.409)$$

Here we have used the contraction rule to write

$$\mathbf{x}^2 = \mathbf{x} \cdot \mathbf{x} = x_1^2 + \dots + x_N^2 \quad (2.410)$$

For complex vectors, we may set  $N = 2m$ , where the dimension of the complex vector is  $m$ , and set the variance of each individual in-phase or quadrature components to  $1/2$  to insure that the variance of each individual complex random variable is 1. This leads specifically to

$$p_X(\mathbf{x}) = \frac{1}{(\sqrt{\pi})^m} e^{-\mathbf{x}^2} \quad (2.411)$$

where

$$\mathbf{X} = \begin{bmatrix} X_{1r} \\ \vdots \\ X_{Nr} \\ X_{1i} \\ \vdots \\ X_{Ni} \end{bmatrix} \quad (2.412)$$

The same form applies to either real or complex random vectors. Moreover, the multivariate form has the same form as the univariate case except for the value of  $N$ .

Now consider an invertible linear transformation of  $\mathbf{X}$ :

$$\mathbf{X} = f(\mathbf{Y}) \quad (2.413)$$

By straightforward substitution we find

$$p_Y(\mathbf{y}) = \frac{\det f}{(\sqrt{2\pi})^N} e^{-\frac{f^2(\mathbf{y})}{2}} \quad (2.414)$$

where

$$f^2(\mathbf{y}) = f(\mathbf{y})f(\mathbf{y}) \quad (2.415)$$

is the geometric product of the vector  $f(\mathbf{y})$  with itself (it is important to observe that  $f(\mathbf{y})$  is a vector).

Let us now apply the geometric algebra formulation to the likelihood ratio for detecting a known signal in Gaussian noise. This problem is well known, and in terms of geometric algebra we find that the likelihood ratio is given by

$$\frac{p_Y(\mathbf{y} - \mathbf{s})}{p_Y(\mathbf{y})} = e^{\frac{f^2(\mathbf{y}) - f^2(\mathbf{y} - \mathbf{s})}{2}} \quad (2.416)$$

Consider now

$$\begin{aligned} f^2(\mathbf{y} - \mathbf{s}) &= f(\mathbf{y} - \mathbf{s})f(\mathbf{y} - \mathbf{s}) \\ &= f^2(\mathbf{y}) - \{f(\mathbf{y})f(\mathbf{s}) + f(\mathbf{s})f(\mathbf{y})\} + f^2(\mathbf{s}) \end{aligned} \quad (2.417)$$

Using our previous results for the geometric product, we see that

$$f(\mathbf{y})f(\mathbf{s}) + f(\mathbf{s})f(\mathbf{y}) = 2f(\mathbf{s}) \cdot f(\mathbf{y}) \quad (2.418)$$

Thus we find for the likelihood ratio

$$\frac{p_Y(\mathbf{y} - \mathbf{s})}{p_Y(\mathbf{y})} = e^{f(\mathbf{s}) \cdot f(\mathbf{y}) - \frac{1}{2}f^2(\mathbf{s})} \quad (2.419)$$

The optimal detector (when  $\mathbf{s}$  is known) is given by the log-likelihood ratio

$$f(\mathbf{s}) \cdot f(\mathbf{y}) \geq T \quad (2.420)$$

where  $T$  is an appropriately chosen threshold. In this formulation the vector  $f(\mathbf{y})$ , which is obtained from the observed data  $\mathbf{y}$ , is projected onto the vector  $f(\mathbf{s})$ , which is obtained from the known vector  $\mathbf{s}$ . As a result, Eq. (2.420) describes a *matched filter* detector.

To see how this form compares with the traditional formulation, let the linear transformation be defined as

$$\mathbf{x} = f(\mathbf{y}) = L\mathbf{y} \quad (2.421)$$

where  $L$  is a nonsingular matrix. Thus we have

$$\mathbf{y} = L^{-1}\mathbf{x} = L^{-1}\mathbf{f}(\mathbf{y}) \quad (2.422)$$

Note that because  $\mathbf{X}$  comprises independent and identically distributed samples with unit variance, we have

$$E[f(\mathbf{y})f(\mathbf{y})^T] = E[\mathbf{x}\mathbf{x}^T] = \mathbf{I} \quad (2.423)$$

Thus the linear transformation  $f(\mathbf{y})$  acts to whiten the data vector  $\mathbf{y}$ . Hence we find

$$E[\mathbf{y}\mathbf{y}^T] = E\left[L^{-1}f(\mathbf{y})f(\mathbf{y})^T L^{-T}\right] = E[L^{-1}L^{-T}] \quad (2.424)$$

From this we see

$$\begin{aligned} R &= L^{-1}L^{-T} \\ R^{-1} &= L^T L \end{aligned}$$

From the geometric product of two vectors it follows that

$$\begin{aligned} f(\mathbf{y})f(\mathbf{y})^T &= f(\mathbf{y}) \cdot f(\mathbf{y}) = (Ly)^T Ly = \mathbf{y}^T L^T Ly \\ &= \mathbf{y}^T R^{-1} \mathbf{y} \end{aligned} \quad (2.425)$$

Similarly we find

$$f^2(\mathbf{s}) = \mathbf{s}^T R^{-1} \mathbf{s} \quad (2.426)$$

$$f(\mathbf{s}) \cdot f(\mathbf{y}) = \mathbf{s}^T R^{-1} \mathbf{y} \quad (2.427)$$

The log-likelihood ratio detector in the traditional formulation is therefore

$$\mathbf{s}^T R^{-1} \mathbf{y} \geq T \quad (2.428)$$

Finally, it is straightforward to show

$$\det f = \frac{1}{|R|^{\frac{1}{2}}} \quad (2.429)$$

The formulation in terms of geometric algebra is precisely equal to the traditional formulation. The advantage of the geometric formulation is that it explicitly shows the role of the whitening transformation—the data is whitened, the signal is whitened, and the inner product is computed. This of course is inherent in the traditional approach but is explicitly shown in the geometric algebra formulation.

Without loss of generality, assume now that the random variables are independently and identically distributed complex random variables with variance equal to 1/2 and write

$$\mathbf{s}_{\text{complex}} = s_r + i s_i \quad (2.430)$$

$$\mathbf{y}_{\text{complex}} = y_r + i y_i \quad (2.431)$$

Associate the following  $2N$ -dimensional vectors with these complex vectors:

$$\mathbf{s} = \begin{bmatrix} s_{1r} \\ \vdots \\ s_{Nr} \\ s_{1i} \\ \vdots \\ s_{Ni} \end{bmatrix} = \begin{bmatrix} \mathbf{s}_r \\ \mathbf{s}_i \end{bmatrix} \quad (2.432)$$

$$\mathbf{y} = \begin{bmatrix} y_{1r} \\ \vdots \\ y_{Nr} \\ y_{1i} \\ \vdots \\ y_{Ni} \end{bmatrix} = \begin{bmatrix} \mathbf{y}_r \\ \mathbf{y}_i \end{bmatrix} \quad (2.433)$$

A straightforward computation yields

$$\mathbf{s} \cdot \mathbf{y} = \sum_{k=1}^m s_k y_{kr} + \sum_{k=1}^m s_k i y_{ki} = \mathbf{s}_r \cdot \mathbf{y}_r + \mathbf{s}_i \cdot \mathbf{y}_i \quad (2.434)$$

Compare the relation in Eq. (2.434) to the complex Hermitian inner product:

$$\begin{aligned} \mathbf{s}_{\text{complex}}^* \cdot \mathbf{y}_{\text{complex}} &= (\mathbf{s}_r - i\mathbf{s}_i) \cdot (\mathbf{y}_r + i\mathbf{y}_i) \\ &= \mathbf{s}_r \cdot \mathbf{y}_r + \mathbf{s}_i \cdot \mathbf{y}_i + i(\mathbf{s}_r \cdot \mathbf{y}_i - \mathbf{s}_i \cdot \mathbf{y}_r) \end{aligned} \quad (2.435)$$

Eqs. (2.434) and (2.435) show

$$\mathbf{s} \cdot \mathbf{y} = \operatorname{Re} [\mathbf{s}_{\text{complex}}^* \cdot \mathbf{y}_{\text{complex}}] \quad (2.436)$$

This shows that Eq. (2.420) gives the correct result for the log-like ratio detector when the data is complex as well.

Again assume without loss of generality that the data is independent and identically distributed and let

$$\mathbf{s}_{\text{complex}} = A e^{i\phi} \mathbf{p}_{\text{complex}} \quad (2.437)$$

where  $\mathbf{p}_{\text{complex}}$  is a known  $N$ -dimensional complex steering vector with  $|\mathbf{p}_{\text{complex}}| = 1$  and  $A e^{i\phi}$  is an unknown complex amplitude. In terms of  $2N$ -dimensional real vectors we have

$$\mathbf{s} = \beta \mathbf{p} \quad (2.438)$$

$$\beta = A e^{i_p \phi} \quad (2.439)$$

where  $\mathbf{p}$  is a known  $2N$ -dimensional real vector obtained from  $\mathbf{p}_{\text{complex}}$  as in the previous section,  $i_p$  is the unit bivector in the  $\mathbf{p}, \mathbf{p}'$ -plane, and  $\beta$  is a spinor in this plane. The likelihood ratio becomes

$$\frac{p_Y(\mathbf{y} - \mathbf{s})}{p_Y(\mathbf{y})} = e^{\frac{y^2 - (\mathbf{y} - \mathbf{s})^2}{2}} = e^{\frac{y^2 - (\mathbf{y} - \beta \mathbf{p})^2}{2}} \quad (2.440)$$

Although the results earlier in this section assumed that the signal  $\mathbf{s}$  was known, in the problem of interest here only  $\mathbf{p}$  is known. The spinor  $\beta$ , which represents the unknown complex amplitude of the signal, is unknown. One common approach to this problem is to choose the unknown complex amplitude (and hence the spinor) so as to maximize the likelihood ratio. The resulting detector is called the generalized likelihood ratio test. To maximize the likelihood ratio (2.440) in the geometric formulation, observe first that the vector  $\mathbf{y}$  has a component in the  $\mathbf{p}, \mathbf{p}'$ -plane and a component orthogonal to this plane:

$$\mathbf{y} = \mathbf{y}_\perp + \mathbf{y}_\parallel \quad (2.441)$$

Thus

$$\mathbf{y} - \beta \mathbf{p} = \mathbf{y}_\perp + \mathbf{y}_\parallel - \beta \mathbf{p} \quad (2.442)$$

The likelihood ratio in Eq. (2.440) will be maximized if the magnitude of  $\mathbf{y} - \beta \mathbf{p}$  is made as small as possible. Because  $\beta$  operates in the  $\mathbf{p}, \mathbf{p}'$ -plane and both  $\mathbf{y}_\parallel$  and  $\mathbf{p}$  are in this plane, this minimization occurs when

$$\mathbf{y}_\parallel - \beta \mathbf{p} = 0 \quad (2.443)$$

Multiplying Eq. (2.439) from the right by  $\mathbf{p}$  immediately yields

$$\beta = \mathbf{y}_\parallel \cdot \mathbf{p} \quad (2.444)$$

This is precisely the relation obtained in the previous section at Eq. (2.403) where we now recognize that  $\mathbf{y}_\parallel$ , which is the component of the data vector  $\mathbf{y}$  in the  $\mathbf{p}, \mathbf{p}'$ -plane, is obtained by projecting  $\mathbf{y}$  onto  $i_p$ :

$$\mathbf{y}_\parallel = (\mathbf{y} \cdot i_p) i_p^\dagger \quad (2.445)$$

In terms of the complex vectors this relation for the spinor becomes

$$\beta = (\mathbf{s}_{\text{complex}}^* \cdot \mathbf{y}_{\text{complex}})^* \quad (2.446)$$

With the relations in Eqs. (2.442) and (2.443) the generalized likelihood ratio becomes

$$\frac{p_Y(\mathbf{y} - \mathbf{s})}{p_Y(\mathbf{y})} = e^{\frac{\mathbf{y}^2 - \mathbf{y}_\perp^2}{2}} \quad (2.447)$$

Observe now that

$$\mathbf{y}^2 - \mathbf{y}_\perp^2 = \mathbf{y}_\parallel^2 \quad (2.448)$$

Thus the logarithm of the generalized likelihood ratio test takes the form

$$\mathbf{y}_\parallel^2 \geq T \quad (2.449)$$

where  $T$  is an appropriately chosen threshold (the 2 appearing in the generalized likelihood ratio has been absorbed into the threshold). The detector structure on the left-hand side of Eq. (2.449) is simply the magnitude (squared) of the projection of the observed data vector onto the subspace of the unit bivector  $i_p$ . In terms of the complex vectors we have

$$\mathbf{y}_\parallel = (\mathbf{s}_{\text{complex}}^* \cdot \mathbf{y}_{\text{complex}})^* \mathbf{p}_{\text{complex}} \quad (2.450)$$

From Eq. (2.450) it follows that:

$$\begin{aligned} \mathbf{y}_\parallel^2 &= (\mathbf{s}_{\text{complex}}^* \cdot \mathbf{y}_{\text{complex}})^* \mathbf{p}_{\text{complex}} \mathbf{p}_{\text{complex}}^* (\mathbf{s}_{\text{complex}}^* \cdot \mathbf{y}_{\text{complex}}) \\ &= |\mathbf{s}_{\text{complex}}^* \cdot \mathbf{y}_{\text{complex}}|^2 \end{aligned} \quad (2.451)$$

Here we have used the assumed normalization  $|\mathbf{p}_{\text{complex}}|^2 = |\mathbf{p}|^2 = 1$ . Thus in traditional complex notation the generalized likelihood ratio test in this simple problem is

$$|\mathbf{s}_{\text{complex}}^* \cdot \mathbf{y}_{\text{complex}}|^2 \geq T \quad (2.452)$$

Compare the detectors in Eqs. (2.436) and (2.452). The detector in Eq. (2.436) assumes both the steering vector  $\mathbf{p}_{\text{complex}}$  and the complex amplitude of the signal are known (in particular it assumes that the phase offset is known because the unknown amplitude  $A$  can be absorbed into the threshold), whereas the detector in Eq. (2.452) assumes only that the steering vector  $\mathbf{p}_{\text{complex}}$  is known. As discussed before with respect to Eq. (2.420), the result in Eq. (2.436) is a matched *filter* detector obtained by projecting the observed  $2N$ -dimensional real data vector  $\mathbf{y}$  onto a specific vector. However, the result in Eq. (2.452) is not a matched filter detector—it is a matched *subspace* detector obtained by projecting the observed data vector into the two-dimensional subspace  $i_p$ . This is necessitated by the lack of knowledge of the phase offset. In formulating Eq. (2.436) it is expected that the signal will be along the vector  $\mathbf{p}_\phi$ , which is a specific vector within the subspace  $i_p$ , whereas in formulating Eq. (2.452) it is expected only that the signal will be in the subspace  $i_p$ . Hence the detector in Eq. (2.452) is matched only to the subspace  $i_p$  rather than to a specific vector within that subspace. See Ref. [54] for a much more in-depth discussion of matched subspace detection.

### 2.3.2.2 A Geometric Approach to Formulating Detectors

The above considerations obtained previously known results in a very straightforward way using geometric algebra and thereby gave them a geometric interpretation. In addition to facilitating certain manipulations and permitting a geometric interpretation of existing results, thinking about problems geometrically can also suggest new approaches. In this section we summarize an approach to the detection problem that has been suggested previously in Ref. [23].

Consider a detection problem in the form of a simple binary hypothesis test:

$$\begin{aligned} H_0 : \mathbf{z} &= \mathbf{n} \\ H_1 : \mathbf{z} &= \mathbf{s} + \mathbf{n} \end{aligned} \quad (2.453)$$

where  $\mathbf{n}$  is a real  $N$ -dimensional noise vector and  $\mathbf{s}$  is a known *real*  $N$ -dimensional signal vector. We address the problem of complex data later. Assume without loss of generality

$$\begin{aligned} E[\mathbf{n}] &= 0 \\ \text{Cov}[\mathbf{n}] &= I \end{aligned} \quad (2.454)$$

where  $I$  is an  $N \times N$  identity matrix.<sup>4</sup>

---

<sup>4</sup>If the noise is correlated, the observed data vector  $\mathbf{z}$  can be whitened and the signal vector  $\mathbf{s}$  can be prewhitened.

Observe first that for any particular decision, we have two vectors available: the known signal vector  $s$  and the observed data vector  $z$ . Taken together, these two vectors define a two-dimensional subspace—i.e., plane—in the  $N$ -dimensional space. The task is to divide this plane into two regions such that if the observed data vector  $z$  falls into Region 0, the signal is declared to be absent, whereas if the observed data vector  $z$  falls into Region 1, the signal is declared to be present. This situation is illustrated in Fig. 2.32.

As shown in Fig. 2.32, the separation of the plane into two regions is specified by a curve in the plane. Each observed  $\hat{z}$  (which is a unit vector in the direction of  $z$ ) gives rise to a different plane and hence a different intersecting curve. Define a coordinate system so that  $\hat{y} = \hat{s}$  is parallel to  $s$ —and thus  $\hat{x}$  is perpendicular to  $s$ —with both  $\hat{y}$  and  $\hat{x}$  lying in the  $\hat{s}, \hat{z}$  plane. In this coordinate system one may write a parametric equation for the separating curve in the  $s, z$  plane as

$$y = f_{s, \hat{z}}(x) \quad (2.455)$$

In general, the function  $f$  is defined for the specific  $\hat{s}, \hat{z}$  plane and hence its form depends on both  $\hat{s}$  and  $\hat{z}$ .

Within this plane, the observed data vector may be written as

$$z = z_x \hat{x} + z_y \hat{y}$$

It is evident from the picture in Fig. 2.32 that a detection occurs whenever

$$z_y \geq f_{s, \hat{z}}(z_x) \quad (2.456)$$

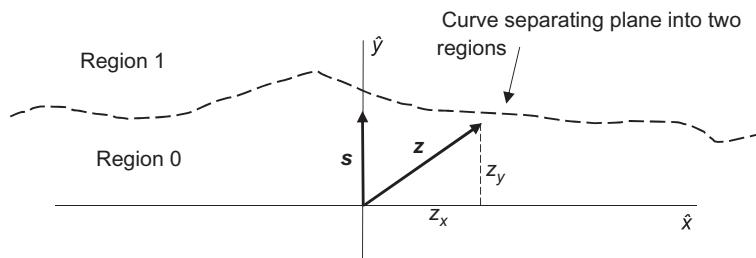
If the component of the observed vector  $z$  in the direction  $\hat{s}$  exceeds some function of the component of  $z$  orthogonal to  $\hat{s}$ , a detection is declared. These two components of the data vector (in the  $s, z$  plane) are easily seen to be

$$z_y = \hat{s} \cdot z \quad (2.457)$$

$$z_x = \pm \sqrt{z \cdot z - (\hat{s} \cdot z)^2} \quad (2.458)$$

Thus simple geometric considerations suggest the following detection structure:

$$\hat{s} \cdot z \geq f_{s, z} \left( \pm \sqrt{z \cdot z - (\hat{s} \cdot z)^2} \right) \quad (2.459)$$



**FIG. 2.32**

Hypothesis testing regions.

Because the form of  $f_{\hat{s},z}(\cdot)$  in general depends on the direction  $\hat{z}$  of the observed vector  $z$ , this detection structure is probably too general to be useful. One simplification is to assume  $f_{s,z}(\cdot)$  has the same *form* for all  $\hat{z}$ . This imposes a symmetry on the problem and leads to

$$\hat{s} \cdot z \geq f_{\hat{s}} \left( \sqrt{z \cdot z - (\hat{s} \cdot z)^2} \right) \quad (2.460)$$

This detection structure, which may be different for each signal vector direction  $\hat{s}$ , is still quite general, but is much less complicated than Eq. (2.459).

To extend these ideas to complex-valued data we now use the geometric interpretation of in-phase and quadrature signals as discussed before. The  $N$ -dimensional complex data vector may be associated with a  $2N$ -dimensional real data vector as has been discussed throughout this work. The question then arises as to what direction to choose for the corresponding signal vector. As has already been discussed in Section 2.3.1, the vector  $d_p$  that is obtained by projecting the observed data onto the unit bivector  $i_p$  has an interpretation as an estimate of the signal. In what follows, assume that the initial signal phase offset is unknown as this is the most common case in radar. Then as discussed in Section 2.3.1, the terms appearing in Eq. (2.460) are replaced by

$$\hat{s} \cdot z \rightarrow |d_p| \quad (2.461)$$

$$z \cdot z - (\hat{s} \cdot z)^2 \rightarrow |d|^2 - |d_p|^2 \quad (2.462)$$

Thus the geometric approach leads to the following detection structure:

$$|d_p| \geq f_s \left( \sqrt{|d|^2 - |d_p|^2} \right) \quad (2.463)$$

Note that

$$d \cdot \frac{d_p}{|d_p|} = |d_p| = |\hat{s}^* \cdot z| \quad (2.464)$$

Thus in going from the real case to the complex case, the  $N$ -dimensional real data vector  $z$  in Eq. (2.460) is replaced by the  $2N$ -dimensional real data vector  $d$ , and the  $N$ -dimensional real steering vector  $\hat{s}$  in Eq. (2.460) is replaced by the  $2N$ -dimensional real estimated signal direction vector  $\frac{d_p}{|d_p|}$ . This further justifies

the interpretation of the magnitude of the Hermitian inner product  $|\hat{s}^* \cdot z|$  as the projection of  $d$  in the direction of  $d_p$ . In terms of the complex vectors, the general detector in Eq. (2.463) may be written

$$|\hat{s}^* \cdot z| \geq f_{\hat{s}} \left( \sqrt{|z|^2 - |\hat{s}^* \cdot z|^2} \right) \quad (2.465)$$

where we have used  $|s| = |p| = 1$  and  $|d| = |z|$ .

The approach now is to choose the function  $f_s(\cdot)$  defining the separating curve directly rather than through the statistics under the two hypotheses. As an example, let

$$f_s(x) = a_s + b_s|x|, \quad a_s, b_s \geq 0 \quad (2.466)$$

This decision region is illustrated in Fig. 2.33.

This choice leads to the following detector structure

$$|\hat{s}^* \cdot z| \geq a_s + b_s \sqrt{|z|^2 - |\hat{s}^* \cdot z|^2} \quad (2.467)$$

When  $b_s=0$  this detector becomes

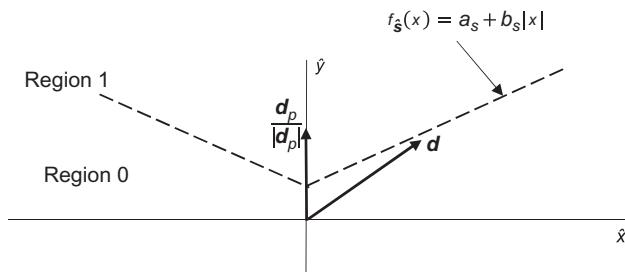
$$|\hat{s}^* \cdot z| \geq a_s \quad (2.468)$$

The detector in Eq. (2.468) is the matched subspace Gaussian detector obtained earlier in this section. When  $a_s=0$  this detector may be rewritten as

$$\frac{|\hat{s}^* \cdot z|}{|z|} \geq \frac{b_s}{\sqrt{b_s^2 + 1}} \quad (2.469)$$

This is the well-known generalized likelihood ratio test [43,46,49,51,53] in which both the unknown complex signal amplitude and the unknown power level of the Gaussian noise are estimated. Thus the detector in Eq. (2.467) is a reasonably simple detector that encompasses both the traditional Gaussian detector and the well-known generalized likelihood ratio test. This result suggests that thinking about the detection problem geometrically may lead to effective detection structures (since in this case the two bounding detectors are undoubtedly effective in appropriate cases).

The detector in Eq. (2.465) is a quite general structure and was obtained from simple geometric reasoning about the detection problem. The availability of such a solution suggests that other geometric insights may possibly be brought to bear on the detection problem to obtain useful detectors. There of course can be no claim to optimality in such an approach, but the geometric approach is not completely arbitrary and may lead to intuitively appealing detectors that can then be checked for acceptable performance. One such example has been presented herein. Other examples may be found in Ref. [23].



**FIG. 2.33**

Separating curve defining a specific detection processor.

### 2.3.3 GEOMETRY OF NULLING DIRECTIONS

#### 2.3.3.1 Linear Processing to Steer Nulls

In this section we examine the application of geometric algebra to linear processing to null certain directions of interest. As in the previous sections, an  $N$ -dimensional complex data vector is associated with a  $2N$ -dimensional real vector and an  $N$ -dimensional complex steering vector is associated with two  $2N$ -dimensional real vectors that together define a two-dimensional subspace. In particular, let

$\mathbf{d}$  =  $2N$ -dimensional real vector associated with an observed  $N$ -dimensional complex data vector  $\mathbf{z}_{\text{complex}}$

$\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_{2k}$  = linearly independent  $2N$ -dimensional real unit vectors associated with complex vectors  $\mathbf{s}_{\text{complex},m}, m=1, \dots, k$  that define directions of desired nulls (with  $k < N$ ).

In this notation the pair  $\mathbf{p}_{2m-1}, \mathbf{p}_{2m}$  is associated with the  $m$ th complex steering vector for  $m=1, \dots, k$  with  $\mathbf{p}_{2m} = \mathbf{p}_{2m-1}'$ . Thus we also have unit bivectors  $i_m = \mathbf{p}_{2m-1}' \mathbf{p}_{2m-1}, m=1, \dots, k$ .

In this problem we have one complex data vector and  $k$  complex steering vectors associated with “directions” that are to be nulled from the data. This means we want to obtain  $\mathbf{z}_{\text{complex},\text{nulled}}$  from  $\mathbf{z}_{\text{complex}}$  such that

$$\mathbf{s}_{\text{complex},m}^* \cdot \mathbf{z}_{\text{complex},\text{nulled}} = 0, \quad m = 1, \dots, k \quad (2.470)$$

We know from previous sections that

$$\mathbf{s}_{\text{complex},m}^* \cdot \mathbf{z}_{\text{complex},\text{nulled}} = \mathbf{p}_m \mathbf{d}_{p_m,\text{nulled}} \quad (2.471)$$

Here,  $\mathbf{d}_{p_m,\text{nulled}}$  is the projection of the vector  $\mathbf{d}_{\text{nulled}}$ , which is associated with the complex vector  $\mathbf{z}_{\text{complex},\text{nulled}}$ , onto the bivector  $i_m, m=1, \dots, k$ . Thus we want

$$\mathbf{d}_{p_m,\text{nulled}} = 0, \quad m = 1, \dots, k \quad (2.472)$$

This states that the vector  $\mathbf{d}_{\text{nulled}}$  is orthogonal to each of the subspaces  $i_m, m=1, \dots, k$ .

Define the subspace

$$A_{2k} = \mathbf{p}_1 \wedge \mathbf{p}_2 \wedge \dots \wedge \mathbf{p}_{2k} = i_1 \wedge i_2 \wedge \dots \wedge i_k \quad (2.473)$$

Because  $\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_{2k}$  are linearly independent, they define a  $2k$ -dimensional subspace of the  $2N$ -dimensional space such that each  $\mathbf{p}_i, i=1, \dots, 2k$  may be expressed as

$$\mathbf{p}_i = \sum_{m=1}^{2k} p_{im} \hat{\mathbf{e}}_m \quad (2.474)$$

where  $\hat{\mathbf{e}}_1, \dots, \hat{\mathbf{e}}_{2k}$  is an orthonormal basis for this  $2k$ -dimensional subspace and  $\hat{\mathbf{e}}_1, \dots, \hat{\mathbf{e}}_{2N}$  is an orthonormal basis for the whole  $2N$ -dimensional space. If we insert Eq. (2.474) into Eq. (2.473) we find

$$A_{2k} \propto \hat{\mathbf{e}}_1 \wedge \hat{\mathbf{e}}_2 \wedge \dots \wedge \hat{\mathbf{e}}_{2k} \quad (2.475)$$

and we will write

$$\begin{aligned} A_{2k} &= |A_{2k}| \hat{\mathbf{e}}_1 \wedge \hat{\mathbf{e}}_2 \wedge \dots \wedge \hat{\mathbf{e}}_{2k} \\ &= |A_{2k}| I_{2k} \end{aligned} \quad (2.476)$$

with

$$|A_{2k}| = |\mathbf{p}_1 \wedge \mathbf{p}_2 \wedge \cdots \wedge \mathbf{p}_{2k}| \quad (2.477)$$

where  $I_{2k}$  is the pseudoscalar for the  $2k$ -dimensional subspace containing  $\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_k$ .

Now examine

$$\mathbf{d} A_{2k} = \mathbf{p} \cdot A_{2k} + \mathbf{d} \wedge A_{2k} \quad (2.478)$$

From this we may write

$$\begin{aligned} \mathbf{d} &= \mathbf{d} A_{2k} A_{2k}^{-1} \\ &= (\mathbf{d} \cdot A_{2k}) A_{2k}^{-1} + (\mathbf{d} \wedge A_{2k}) A_{2k}^{-1}. \end{aligned} \quad (2.479)$$

From Eq. (2.476) it follows that:

$$A_{2k}^{-1} = \frac{I_{2k}^{-1}}{|A_{2k}|} = \frac{(-1)^{\frac{2k(2k-1)}{2}}}{|A_{2k}|} I_k = (-1)^{\frac{2k(2k-1)}{2}} \frac{A_{2k}}{|A_{2k}|^2}. \quad (2.480)$$

We now find from Eq. (2.479)

$$\begin{aligned} \mathbf{d} &= (-1)^{\frac{2k(2k-1)}{2}} (\mathbf{d} \cdot A_{2k}) \frac{A_{2k}}{|A_{2k}|^2} + (-1)^{\frac{2k(2k-1)}{2}} (\mathbf{d} \wedge A_{2k}) \frac{A_{2k}}{|A_{2k}|^2} \\ &= (\mathbf{d} \cdot A_{2k}) \frac{A_{2k}^\dagger}{|A_{2k}|^2} + (\mathbf{d} \wedge A_{2k}) \frac{A_{2k}^\dagger}{|A_{2k}|^2} \end{aligned} \quad (2.481)$$

Eq. (2.481) decomposes  $\mathbf{d}$  into its projection onto the subspace  $A_{2k}$  determined by the nulling vectors  $\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_{2k}$  (or equivalently the nulling bivectors  $i_1, i_2, \dots, i_k$ ) and the part that is not in the subspace  $A_{2k}$  (i.e., the part that is orthogonal to it). These two components of  $\mathbf{d}$  are called the projection and rejection:

$$(\mathbf{d} \cdot A_{2k}) \frac{A_{2k}^\dagger}{|A_{2k}|^2} = \text{Projection of } \mathbf{d} \text{ onto } A_{2k} \quad (2.482)$$

$$(\mathbf{d} \wedge A_{2k}) \frac{A_{2k}^\dagger}{|A_{2k}|^2} = \text{Rejection of } \mathbf{d} \text{ from } A_{2k} \quad (2.483)$$

Since the rejection of  $\mathbf{d}$  from  $A_{2k}$  is a vector that is orthogonal to the subspace  $A_{2k}$  which comprises the subspaces  $i_m, m=1, \dots, k$ , it follows that the desired vector  $\mathbf{d}_{\text{nulled}}$  is given by the rejection of  $\mathbf{d}$  from  $A_{2k}$ :

$$\mathbf{d}_{\text{nulled}} = (\mathbf{d} \wedge A_{2k}) \frac{A_{2k}^\dagger}{|A_{2k}|^2} \quad (2.484)$$

With this choice, projecting  $\mathbf{d}_{\text{nulled}}$  onto any of the subspaces  $i_m, m=1, \dots, k$  yields  $\mathbf{d}_{\mathbf{p}_m, \text{nulled}} = 0$  as desired.

To examine this result further, now define

$$\mathbf{d}_{\perp, 2k} = (\mathbf{d} \wedge A_{2k}) \frac{A_{2k}^\dagger}{|A_{2k}|^2} = \mathbf{d} - (\mathbf{d} \cdot A_{2k}) \frac{A_{2k}^\dagger}{|A_{2k}|^2} \quad (2.485)$$

With Eq. (2.476) this becomes

$$\begin{aligned}\mathbf{d}_{\perp,2k} &= \mathbf{d} - (\mathbf{d} \cdot I_{2k}) I_{2k}^\dagger \\ &= (\mathbf{d} \wedge I_{2k}) I_{2k}^\dagger\end{aligned}\quad (2.486)$$

Let us now express  $\mathbf{d}$  in terms of our basis as

$$\mathbf{d} = \alpha_1 \hat{\mathbf{e}}_1 + \alpha_2 \hat{\mathbf{e}}_2 + \alpha_3 \hat{\mathbf{e}}_3 + \cdots + \alpha_{2N} \hat{\mathbf{e}}_{2N}. \quad (2.487)$$

With this representation we therefore have

$$\mathbf{d}_{\perp,2k} = \left( \sum_{m=1}^{2N} \alpha_m \hat{\mathbf{e}}_m \wedge I_{2k} \right) I_{2k}^\dagger \quad (2.488)$$

Examine

$$\hat{\mathbf{e}}_m \wedge I_{2k} = \hat{\mathbf{e}}_m \wedge \hat{\mathbf{e}}_1 \dots \hat{\mathbf{e}}_1. \quad (2.489)$$

When  $m = 1, \dots, 2k$ , then  $\hat{\mathbf{e}}_m$  is parallel to one of the vectors defining  $I_{2k}$  but when  $m = 2k+1, \dots, N$   $\hat{\mathbf{e}}_m$  is perpendicular to each of the vectors defining  $I_{2k}$ . It follows then that

$$\hat{\mathbf{e}}_m \wedge I_{2k} = 0, \quad m = 1, \dots, 2k \quad (2.490)$$

and

$$\hat{\mathbf{e}}_m \wedge I_{2k} = \hat{\mathbf{e}}_m \hat{\mathbf{e}}_1 \dots \hat{\mathbf{e}}_{2k} = \hat{\mathbf{e}}_m I_{2k}, \quad m = 2k+1, \dots, 2N. \quad (2.491)$$

Thus

$$\mathbf{d}_{\perp,2k} = \left( \sum_{m=2k+1}^{2N} \alpha_m \hat{\mathbf{e}}_m \wedge I_{2k} \right) I_{2k}^\dagger = \sum_{m=k+1}^N \alpha_m \hat{\mathbf{e}}_m I_k I_k^\dagger \quad (2.492)$$

But

$$I_k I_k^\dagger = 1. \quad (2.493)$$

Therefore

$$\mathbf{d}_{\perp,2k} = \sum_{m=2k+1}^N \alpha_m \hat{\mathbf{e}}_m \quad (2.494)$$

Eq. (2.494) shows that  $\mathbf{d}_{\perp,2k}$  is the projection of the vector  $\mathbf{d}$  onto the space that is orthogonal to the subspace  $A_{2k}$  containing the null directions  $\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_{2k}$ .

These considerations show that *rejecting* a steering vector  $\mathbf{d}$  from a subspace  $A_{2k}$  can be used to null directions. In particular nulling a direction results from projecting the data onto a subspace that is orthogonal to the nulled direction. This kind of processing is implemented quite often in radar processing. As an example, in the next section we explore the application of this kind of geometric approach to nulling portions of a DFT response.

### 2.3.3.2 Geometric Approach to Designing a Notch Filter

Examine the following DFT frequency response:

$$\begin{aligned} X_k(f) &= \frac{1}{N} \sum_{l=0}^{N-1} e^{-i2\pi \frac{lk}{N}} e^{i2\pi lf} \\ &= \frac{1}{N} \sum_{l=0}^{N-1} \left\{ \cos \left( 2\pi l \left[ f - \frac{k}{N} \right] \right) + i \sin \left( 2\pi l \left[ f - \frac{k}{N} \right] \right) \right\} \end{aligned} \quad (2.495)$$

This represents the response of DFT filter  $k, k=0, \dots, N-1$  to an input sinusoidal signal with (normalized) frequency  $f \in [0, 1]$ . Define two orthonormal  $2N$ -dimensional vectors:

$$\boldsymbol{\varphi}_f = \frac{1}{\sqrt{N}} \begin{bmatrix} \cos(2\pi 0f) \\ \cos(2\pi 1f) \\ \vdots \\ \cos(2\pi(N-1)f) \\ \sin(2\pi 0f) \\ \sin(2\pi 1f) \\ \vdots \\ \sin(2\pi(N-1)f) \end{bmatrix}; \quad \boldsymbol{\varphi}'_f = \frac{1}{\sqrt{N}} \begin{bmatrix} -\sin(2\pi 0f) \\ -\sin(2\pi 1f) \\ \vdots \\ -\sin(2\pi(N-1)f) \\ \cos(2\pi 0f) \\ \cos(2\pi 1f) \\ \vdots \\ \cos(2\pi(N-1)f) \end{bmatrix} \quad (2.496)$$

Examine

$$\begin{aligned} \boldsymbol{\varphi}_{f_1} \cdot \boldsymbol{\varphi}_{f_2} &= \boldsymbol{\varphi}'_{f_1} \cdot \boldsymbol{\varphi}'_{f_2} = \frac{1}{N} \sum_{l=0}^{N-1} \{ \cos(2\pi lf_1) \cos(2\pi lf_2) + \sin(2\pi lf_1) \sin(2\pi lf_2) \} \\ &= \frac{1}{N} \sum_{l=0}^{N-1} \cos(2\pi l[f_1 - f_2]) \end{aligned} \quad (2.497)$$

Similarly

$$\begin{aligned} \boldsymbol{\varphi}_{f_1} \cdot \boldsymbol{\varphi}'_{f_2} &= -\boldsymbol{\varphi}'_{f_1} \cdot \boldsymbol{\varphi}_{f_2} \\ &= \frac{1}{N} \sum_{l=0}^{N-1} \{ -\cos(2\pi lf_1) \sin(2\pi lf_2) + \sin(2\pi lf_1) \cos(2\pi lf_2) \} \\ &= \frac{1}{N} \sum_{l=0}^{N-1} \sin(2\pi l[f_1 - f_2]) \end{aligned} \quad (2.498)$$

Setting  $f_1 = f_2$  in these expressions establishes the orthonormality of these vectors. From these relations the DFT response may be reexpressed in terms of the vectors  $\boldsymbol{\varphi}_{\frac{k}{N}}$  and  $\boldsymbol{\varphi}'_{\frac{k}{N}}$  as

$$X_k(f) = X_k(\boldsymbol{\varphi}_f) = \frac{1}{N} \sum_{l=0}^{N-1} e^{-i2\pi \frac{lk}{N}} e^{i2\pi lf} = \boldsymbol{\varphi}_f \cdot \boldsymbol{\varphi}_{\frac{k}{N}} + i \boldsymbol{\varphi}_f \cdot \boldsymbol{\varphi}'_{\frac{k}{N}} \quad (2.499)$$

More generally, if  $c_0, c_1, \dots, c_{N-1}$  are complex samples of an input signal to the DFT, then the DFT response to this input sequence may be written as

$$X_k(\mathbf{d}) = \frac{1}{N} \sum_{l=0}^{N-1} e^{-i2\pi \frac{lk}{N}} c_l = \mathbf{d} \cdot \boldsymbol{\varphi}_{\frac{k}{N}} + i \mathbf{d} \cdot \boldsymbol{\varphi}'_{\frac{k}{N}} \quad (2.500)$$

where the  $2N$ -dimensional real vector  $\mathbf{d}$  is formed from the  $N$  complex samples  $\{c_0, c_1, \dots, c_{N-1}\}$  by assembling the real parts and the imaginary parts into two orthogonal vectors as follows:

$$\mathbf{d} = \begin{bmatrix} c_{0,real} \\ c_{1,real} \\ \vdots \\ c_{N-1,real} \\ c_{0,imag} \\ c_{1,imag} \\ \vdots \\ c_{N-1,imag} \end{bmatrix}; \quad \mathbf{d}' = \begin{bmatrix} -c_{0,imag} \\ -c_{1,imag} \\ \vdots \\ -c_{N-1,imag} \\ c_{0,real} \\ c_{1,real} \\ \vdots \\ c_{N-1,real} \end{bmatrix} \quad (2.501)$$

Observe now that the previous considerations imply

$$\mathbf{d} \cdot \boldsymbol{\varphi}_{\frac{k}{N}} = \mathbf{d}' \cdot \boldsymbol{\varphi}'_{\frac{k}{N}} \quad (2.502)$$

$$\mathbf{d} \cdot \boldsymbol{\varphi}'_{\frac{k}{N}} = -\mathbf{d}' \cdot \boldsymbol{\varphi}_{\frac{k}{N}} \quad (2.503)$$

Therefore

$$X_k(\mathbf{d}) = \mathbf{d}' \cdot \boldsymbol{\varphi}'_{\frac{k}{N}} - i \mathbf{d}' \cdot \boldsymbol{\varphi}_{\frac{k}{N}} = -i X_k(\mathbf{d}') \quad (2.504)$$

This relation expresses the fact that the vector  $\mathbf{d}$  is obtained from the sequence  $\{c_0, c_1, \dots, c_{N-1}\}$  whereas the vector  $\mathbf{d}'$  is obtained from the sequence  $\{-ic_0, -ic_1, \dots, -ic_{N-1}\}$ , and hence the DFT responses to these respective input sequences are related by the factor  $-i$ .

The interest in this section is now to modify the implementation of the DFT so as to null the response at selected frequencies. We begin by defining a unit bivector formed from the orthogonal vectors  $\boldsymbol{\varphi}_{\frac{k}{N}}$  and  $\boldsymbol{\varphi}'_{\frac{k}{N}}$  that define the DFT filter:

$$i_k = \boldsymbol{\varphi}'_{\frac{k}{N}} \wedge \boldsymbol{\varphi}_{\frac{k}{N}} = \boldsymbol{\varphi}'_{\frac{k}{N}} \boldsymbol{\varphi}_{\frac{k}{N}} \quad (2.505)$$

As has been discussed in a previous section, the orthogonal projection of an arbitrary vector  $\mathbf{d}$  onto the bivector  $i_k$  is given by

$$\mathbf{d}_{\parallel} = (\mathbf{d} \cdot i_k) i_k^{\dagger} = (\mathbf{d} \cdot \boldsymbol{\varphi}_{\frac{k}{N}}) \boldsymbol{\varphi}_{\frac{k}{N}} + (\mathbf{d} \cdot \boldsymbol{\varphi}'_{\frac{k}{N}}) \boldsymbol{\varphi}'_{\frac{k}{N}} \quad (2.506)$$

In this expression,  $\mathbf{d}_{\parallel}$  is the component of the vector  $\mathbf{d}$  that lies in the two-dimensional subspace determined by  $i_k$ . Use of Eq. (2.500) and replacing the imaginary unit  $i$  with the unit bivector  $i_k$  now shows that the DFT response can be reexpressed in terms of geometric algebra as:

$$\begin{aligned}
X_k(\mathbf{d}) &= \left( \mathbf{d} \cdot \boldsymbol{\varphi}_{\frac{k}{N}} \right) + i_k \left( \mathbf{d} \cdot \boldsymbol{\varphi}'_{\frac{k}{N}} \right) \\
&= \left[ \left( \mathbf{d} \cdot \boldsymbol{\varphi}_{\frac{k}{N}} \right) \boldsymbol{\varphi}_{\frac{k}{N}} \boldsymbol{\varphi}_{\frac{k}{N}} + \left( \mathbf{d} \cdot \boldsymbol{\varphi}'_{\frac{k}{N}} \right) \boldsymbol{\varphi}'_{\frac{k}{N}} \boldsymbol{\varphi}_{\frac{k}{N}} \right] \\
&= \left[ \left( \mathbf{d} \cdot \boldsymbol{\varphi}_{\frac{k}{N}} \right) \boldsymbol{\varphi}_{\frac{k}{N}} + \left( \mathbf{d} \cdot \boldsymbol{\varphi}'_{\frac{k}{N}} \right) \boldsymbol{\varphi}'_{\frac{k}{N}} \right] \boldsymbol{\varphi}_{\frac{k}{N}} \\
&= \left[ (\mathbf{d} \cdot i_k) i_k^\dagger \right] \boldsymbol{\varphi}_{\frac{k}{N}}
\end{aligned} \tag{2.507}$$

This filter is obtained by projecting the data vector  $\mathbf{d}$  onto the bivector  $i_k$  and then forming the geometric product of the projected vector with the vector  $\boldsymbol{\varphi}_{\frac{k}{N}}$ , which lies in the  $i_k$  subspace.

From the formulation in Eq. (2.507), it evidently follows that replacing the unit bivector onto which the data is projected leads to a new filter. In particular, define a new filter in this geometric algebra formulation as

$$X_{w,k}(\mathbf{d}) = \left[ (\mathbf{d} \cdot i_{w,k}) i_{w,k}^\dagger \right] \hat{\mathbf{w}}_k \tag{2.508}$$

The filter in Eq. (2.508) is obtained by projecting the data vector  $\mathbf{d}$  onto a new plane defined by a unit bivector  $i_{w,k}$  where  $\hat{\mathbf{w}}_k$  is a unit vector in the  $i_{w,k}$ -plane.

This filter response may reexpressed as

$$\begin{aligned}
X_{w,k}(\mathbf{d}) &= [(\mathbf{d} \cdot \hat{\mathbf{w}}_k) \hat{\mathbf{w}}_k + (\mathbf{d} \cdot \hat{\mathbf{w}}_{k,\perp}) \hat{\mathbf{w}}_{k,\perp}] \hat{\mathbf{w}}_k \\
&= (\mathbf{d} \cdot \hat{\mathbf{w}}_k) + i_{w,k} (\mathbf{d} \cdot \hat{\mathbf{w}}_{k,\perp})
\end{aligned} \tag{2.509}$$

where

$$i_{w,k} = \hat{\mathbf{w}}_{k,\perp} \wedge \hat{\mathbf{w}}_k = \hat{\mathbf{w}}_{k,\perp} \hat{\mathbf{w}}_k$$

where  $\hat{\mathbf{w}}_k$  and  $\hat{\mathbf{w}}_{k,\perp}$  are orthonormal vectors. Therefore  $i_{w,k}$  is a unit bivector. In essence, the new filter is obtained by replacing the orthonormal vectors  $\boldsymbol{\varphi}_{\frac{k}{N}}$  and  $\boldsymbol{\varphi}'_{\frac{k}{N}}$  with the orthonormal vectors  $\hat{\mathbf{w}}_k$  and  $\hat{\mathbf{w}}_{k,\perp}$ .

The vectors  $\hat{\mathbf{w}}_k$  and  $\hat{\mathbf{w}}_{k,\perp}$  will be chosen so that  $\hat{\mathbf{w}}_k$  and  $i_{w,k}$  are as “close as possible” to  $\boldsymbol{\varphi}_{\frac{k}{N}}$  and  $i_k$ , respectively, while also satisfying certain nulling constraints.

In particular, let  $\boldsymbol{\varphi}_{f_m}, m=1, \dots, M$  be a linearly independent set of vectors representing sinusoidal signals at frequencies  $f_1, f_2, \dots, f_M$  that are to be nulled. Then we want

$$X_{w,k}(\boldsymbol{\varphi}_{f_m}) = \left[ (\boldsymbol{\varphi}_{f_m} \cdot i_{w,k}) i_{w,k}^\dagger \right] \hat{\mathbf{w}}_k = 0 \tag{2.510}$$

This relation in turn implies

$$\boldsymbol{\varphi}_{f_m} \cdot i_{w,k} = \left( \boldsymbol{\varphi}_{f_m} \cdot \boldsymbol{\varphi}'_{\frac{k}{N}} \right) \boldsymbol{\varphi}_{\frac{k}{N}} - \left( \boldsymbol{\varphi}_{f_m} \cdot \boldsymbol{\varphi}_{\frac{k}{N}} \right) \boldsymbol{\varphi}'_{\frac{k}{N}} = 0 \tag{2.511}$$

Let  $I_{2M}$  be the unit pseudoscalar associated with the  $2M$ -dimensional subspace defined by the bivectors  $\boldsymbol{\varphi}_{f_m}' \wedge \boldsymbol{\varphi}_{f_m} / |\boldsymbol{\varphi}_{f_m}' \wedge \boldsymbol{\varphi}_{f_m}| = \boldsymbol{\varphi}_{f_{m,orth}}' \boldsymbol{\varphi}_{f_{m,orth}}, m=1, \dots, M$ . In this expression,  $\boldsymbol{\varphi}_{f_{m,orth}}$  and  $\boldsymbol{\varphi}_{f_{m,orth}}', m=1, \dots, M$  are orthonormalized versions (in the plane) of  $\boldsymbol{\varphi}_{f_m}$  and  $\boldsymbol{\varphi}_{f_m}'$ , respectively. Then the constraints require  $i_{w,k}$  to be

orthogonal to  $I_{2M}$ . In other words, when the data is orthogonally projected into the  $i_{w,k}$  subspace, no part of the projected vector can be in the  $I_{2M}$  subspace. Therefore

$$\begin{aligned} i_{w,k} &= \frac{(i_k \wedge I_{2M})I_{2M}^\dagger}{\|(i_k \wedge I_{2M})I_{2M}^\dagger\|} \\ &= \frac{i_k - (i_k \cdot I_{2M})I_{2M}^\dagger}{\|i_k - (i_k \cdot I_{2M})I_{2M}^\dagger\|} \end{aligned} \quad (2.512)$$

Now define two vectors  $\mathbf{w}_k, \mathbf{w}'_k$  as the orthogonal rejections of the DFT vectors  $\boldsymbol{\varphi}_{\frac{k}{N}}, \boldsymbol{\varphi}'_{\frac{k}{N}}$  from the constraint subspace  $I_{2M}$ :

$$\mathbf{w}_k = \left( \boldsymbol{\varphi}_{\frac{k}{N}} \wedge I_{2M} \right) I_{2M}^\dagger = \boldsymbol{\varphi}_{\frac{k}{N}} - \left( \boldsymbol{\varphi}_{\frac{k}{N}} \cdot I_{2M} \right) I_{2M}^\dagger \quad (2.513)$$

$$\mathbf{w}'_k = \left( \boldsymbol{\varphi}'_{\frac{k}{N}} \wedge I_{2M} \right) I_{2M}^\dagger = \boldsymbol{\varphi}'_{\frac{k}{N}} - \left( \boldsymbol{\varphi}'_{\frac{k}{N}} \cdot I_{2M} \right) I_{2M}^\dagger \quad (2.514)$$

From this it follows that:

$$\begin{aligned} \mathbf{w}'_k \wedge \mathbf{w}_k &= \left( \boldsymbol{\varphi}'_{\frac{k}{N}} \wedge I_{2M} \right) I_{2M}^\dagger \wedge \left( \boldsymbol{\varphi}_{\frac{k}{N}} \wedge I_{2M} \right) I_{2M}^\dagger \\ &= \left( \left( \boldsymbol{\varphi}'_{\frac{k}{N}} \wedge \boldsymbol{\varphi}_{\frac{k}{N}} \right) \wedge I_{2M} \right) I_{2M}^\dagger \\ &= (i_k \wedge I_{2M})_{2M}^\dagger = i_k - (i_k \cdot I_{2M})I_{2M}^\dagger \end{aligned} \quad (2.515)$$

Therefore from Eq. (2.512) the unit bivector  $i_{w,k}$  is related to the vectors  $\mathbf{w}_k, \mathbf{w}'_k$  as

$$i_{w,k} = \frac{\mathbf{w}'_k \wedge \mathbf{w}_k}{\|\mathbf{w}'_k \wedge \mathbf{w}_k\|} \quad (2.516)$$

Note that  $\mathbf{w}_k$  and  $\mathbf{w}'_k$  are not necessarily orthogonal, so the bivector  $\mathbf{w}'_k \wedge \mathbf{w}_k$  must be normalized to obtain a unit bivector.

Explicitly, the vectors  $\mathbf{w}_k$  and  $\mathbf{w}'_k$  are given by

$$\mathbf{w}_k = \boldsymbol{\varphi}_{\frac{k}{N}} - \sum_{m=1}^M \left( \boldsymbol{\varphi}_{\frac{k}{N}} \cdot \boldsymbol{\varphi}_{f_m, \text{orth}} \right) \boldsymbol{\varphi}_{f_m, \text{orth}} - \sum_{m=1}^M \left( \boldsymbol{\varphi}_{\frac{k}{N}} \cdot \boldsymbol{\varphi}'_{f_m} \right) \boldsymbol{\varphi}'_{f_m, \text{orth}} \quad (2.517)$$

$$\mathbf{w}'_k = \boldsymbol{\varphi}'_{\frac{k}{N}} - \sum_{m=1}^M \left( \boldsymbol{\varphi}'_{\frac{k}{N}} \cdot \boldsymbol{\varphi}_{f_m, \text{orth}} \right) \boldsymbol{\varphi}_{f_m, \text{orth}} - \sum_{m=1}^M \left( \boldsymbol{\varphi}'_{\frac{k}{N}} \cdot \boldsymbol{\varphi}'_{f_m, \text{orth}} \right) \boldsymbol{\varphi}'_{f_m, \text{orth}} \quad (2.518)$$

The unit vectors  $\hat{\mathbf{w}}_k, \hat{\mathbf{w}}_{k,\perp}$  are now obtained as

$$\hat{\mathbf{w}}_k = \frac{\mathbf{w}_k}{\|\mathbf{w}_k\|} \quad (2.519)$$

$$\mathbf{w}_{k,\perp} = \mathbf{w}'_k - (\mathbf{w}'_k \cdot \hat{\mathbf{w}}_k) \hat{\mathbf{w}}_k \quad (2.520)$$

$$\hat{\mathbf{w}}_{k,\perp} = \frac{\mathbf{w}_{k,\perp}}{\|\mathbf{w}_{k,\perp}\|} \quad (2.521)$$

It should be noted that in general

$$\hat{\mathbf{w}}_{k,\perp} \neq \hat{\mathbf{w}}'_k \quad (2.522)$$

Thus the new filter is *not* obtained simply by subtracting the DFT response at the nulling frequencies from the DFT response itself.

Now examine the expression  $\mathbf{d} \cdot i_{w,k}$  appearing in the filter. From Eq. (2.512) we may write

$$\mathbf{d} \cdot i_{w,k} = \frac{\mathbf{d} \cdot i_k - \mathbf{d} \cdot (i_k \cdot I_{2M}) I_{2M}^\dagger}{|i_k - (i_k \cdot I_{2M}) I_{2M}^\dagger|} \quad (2.523)$$

Let  $\mathbf{d} = \boldsymbol{\varphi}_f \in I_{2M}$  be a vector within the constraint subspace. Then

$$\boldsymbol{\varphi}_f \cdot i_k = \boldsymbol{\varphi}_f \cdot (i_k \cdot I_{2M}) I_{2M}^\dagger \quad (2.524)$$

This expression merely reflects the fact that  $\boldsymbol{\varphi}_f$  is wholly within  $I_{2M}$  and therefore can only project into the part of  $i_k$  that is also in  $I_{2M}$ . Therefore for any frequency for which  $\boldsymbol{\varphi}_f \in I_m$ , it follows that:

$$\boldsymbol{\varphi}_f \cdot i_{w,k} = 0 \quad (2.525)$$

$$X_{w,k}(\boldsymbol{\varphi}_f) = 0 \quad (2.526)$$

From the relation  $X_{w,k}(\boldsymbol{\varphi}_f) = -i_{w,k} X_{w,k}(\boldsymbol{\varphi}'_f)$  it also follows that in this case

$$X_{w,k}(\boldsymbol{\varphi}'_f) = 0 \quad (2.527)$$

Thus the constraints—i.e., the requirement that certain frequencies be nulled—are satisfied.

Examine the expression for  $w_k$  in Eq. (2.513):

$$w_k = \boldsymbol{\varphi}_{\frac{k}{N}} - (\boldsymbol{\varphi}_{\frac{k}{N}} \cdot I_{2M}) I_{2M}^\dagger \quad (2.528)$$

In this expression  $I_{2M}$  is the unit pseudoscalar associated with the subspace defined by the constraint vectors  $\boldsymbol{\varphi}_{f_m}, m=1, \dots, M$ . Thus  $(\boldsymbol{\varphi}_{\frac{k}{N}} \cdot I_{2M}) I_{2M}^\dagger$  is the projection of the DFT vector  $\boldsymbol{\varphi}_{\frac{k}{N}}$  into the subspace spanned by the constraint vectors. In a traditional complex matrix formulation, this relation may be expressed as follows. Define a complex matrix as

$$C = \begin{bmatrix} e^{j\omega_1} & e^{j2\omega_1} & \dots & e^{jN\omega_1} \\ e^{j\omega_2} & e^{j2\omega_2} & \dots & e^{jN\omega_2} \\ \vdots & 0 & \ddots & 0 \\ e^{j\omega_m} & e^{j2\omega_m} & 0 & e^{jN\omega_m} \end{bmatrix} \quad (2.529)$$

where

$$\omega_m = 2\pi f_m, \quad m = 1, \dots, M \quad (2.530)$$

In terms of this matrix, the constraints that null the frequencies  $f_m, m=1, \dots, M$  are expressed as

$$C \hat{\mathbf{w}}_{complex,k} = 0 \quad (2.531)$$

where  $\hat{\mathbf{w}}_{\text{complex},k}$  is the complex vector associated with  $\hat{\mathbf{w}}_k$ . Let  $\boldsymbol{\varphi}_{\text{complex},\frac{k}{N}}$ , which defines the complex DFT weights, be the complex vector associated with  $\boldsymbol{\varphi}_k$ . Then the projection of the complex vector  $\boldsymbol{\varphi}_{\text{complex},\frac{k}{N}}$  onto the row space of the matrix  $C$  corresponds to the projection of  $\boldsymbol{\varphi}_k$  into  $I_{2M}$ :

$$\left(\boldsymbol{\varphi}_k \cdot I_{2M}\right) I_{2M}^\dagger \leftrightarrow C^H [CC^H]^{-1} C \boldsymbol{\varphi}_{\text{complex},\frac{k}{N}} \quad (2.532)$$

Therefore in a traditional matrix formulation the modified DFT weight is given by

$$\hat{\mathbf{w}}_{\text{complex},k} = \boldsymbol{\varphi}_{\text{complex},\frac{k}{N}} - C^H [CC^H]^{-1} C \boldsymbol{\varphi}_{\text{complex},\frac{k}{N}} \quad (2.533)$$

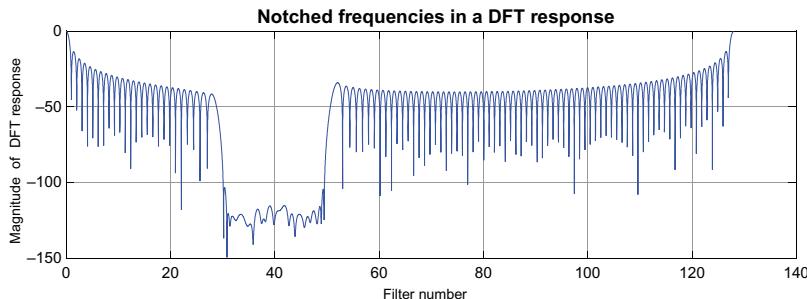
This latter formulation may be obtained directly by solving a constrained least-squares problem.

This geometric approach described herein appears to be an example of projection methods that are applied to adaptive processing in Refs. [45,48,57]. Fig. 2.34 shows an example of nulling a portion of a DFT response in this way. In this example, the DFT has length 128 and the nulling frequencies were chosen to eliminate the response between filter numbers 35 and 50. As shown, the frequency response for filters 35–50 is greatly reduced while the response at other filter numbers is essentially the DFT response. As discussed further later, in implementing this approach one must choose the frequencies to be nulled in somewhat of a careful manner or the approach runs into difficulties.

### 2.3.3.3 Choosing the Frequencies That Define the Constraint Subspace

To determine how to choose the nulling frequencies and thus the constraint subspace, it is of interest to understand the geometric relationships between the vectors  $\boldsymbol{\varphi}_{f_1}, \boldsymbol{\varphi}_{f_2}, \boldsymbol{\varphi}'_{f_1}$  and  $\boldsymbol{\varphi}'_{f_2}$  at different frequencies. First, let  $f_1 = f_2 = f$ . It follows from the relations given previously that

$$\boldsymbol{\varphi}_f \cdot \boldsymbol{\varphi}_f = \boldsymbol{\varphi}'_f \cdot \boldsymbol{\varphi}'_f = \frac{1}{N} \sum_{l=0}^{N-1} \cos(0) = 1 \quad (2.534)$$



**FIG. 2.34**

Notched DFT response.

$$\boldsymbol{\varphi}_f \cdot \boldsymbol{\varphi}'_f = -\boldsymbol{\varphi}'_f \cdot \boldsymbol{\varphi}_f = \frac{1}{N} \sum_{l=0}^{N-1} \sin(0) = 0 \quad (2.535)$$

This shows that for a given frequency,  $\boldsymbol{\varphi}_f$  and  $\boldsymbol{\varphi}'_f$  have unit magnitude and  $\boldsymbol{\varphi}_f$  is orthogonal to  $\boldsymbol{\varphi}'_f$ .

Assume now that  $f_1 \neq f_2$ . Then the first dot product yields:

$$\boldsymbol{\varphi}_{f_1} \cdot \boldsymbol{\varphi}_{f_2} = \boldsymbol{\varphi}'_{f_1} \cdot \boldsymbol{\varphi}'_{f_2} = \frac{1 - \cos(2\pi N[f_1 - f_2])}{2N} + \frac{\sin(2\pi N[f_1 - f_2]) \sin(2\pi [f_1 - f_2])}{2N \{1 - \cos(2\pi [f_1 - f_2])\}} \quad (2.536)$$

The second dot product yields:

$$\boldsymbol{\varphi}_{f_1} \cdot \boldsymbol{\varphi}'_{f_2} = -\boldsymbol{\varphi}'_{f_1} \cdot \boldsymbol{\varphi}_{f_2} = \frac{-\sin(2\pi N[f_1 - f_2])}{2N} + \frac{\{1 - \cos(2\pi N[f_1 - f_2])\} \sin(2\pi [f_1 - f_2])}{2N \{1 - \cos(2\pi [f_1 - f_2])\}} \quad (2.537)$$

These relations are illustrated in Figs. 2.35–2.38 for  $N = 32$ . The  $x$ -axis gives the frequency difference  $f_1 - f_2$ .

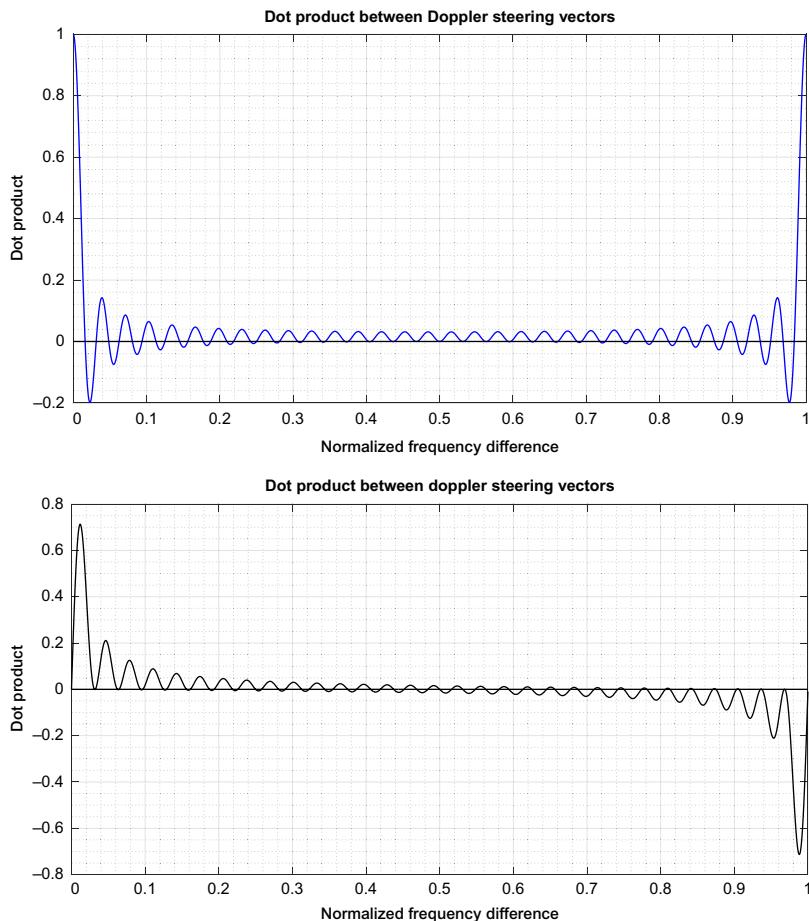
In Figs. 2.36–2.38 these curves are overlaid and expanded with asterisks placed at  $\Delta f = \frac{k}{N}, k = 1, \dots, 32$ . These points represent the frequency differences between the different filters of a DFT.

It is apparent that the local peaks of the two dot products do not generally coincide. Moreover, the relation between them varies with the region of frequency difference of interest. For example, Fig. 2.39 shows plots for the regions of frequency difference [0.05, 0.3] and [0.6, 0.95] for the case  $N = 32$ .

In Fig. 2.39, when one of the dot products has a local peak the other dot product is close to zero. In addition, the second dot product (black curve) essentially has opposite signs in these two plots.

This kind of decomposition of the DFT response may be used to explore how to choose the frequencies that define the constraint subspace. For example, it is apparent that the DFT filter frequencies should not be included in the constraint subspace because those frequencies are already nulled by the DFT. It appears in general that the frequencies should be chosen in pairs corresponding to the local peaks of each of the two dot products in the frequency region to be nulled. For example, in Fig. 2.40 showing for  $N = 32$  the difference frequency region between 0.4 and 0.5, the local peaks of the first dot product occur at approximately 0.404, 0.42, 0.436, 0.452, 0.468, 0.484, 0.5 whereas the local peaks of the second dot product occur at approximately 0.412, 0.428, 0.444, 0.46, 0.476, 0.492.

For both dot products, the spacing between the respective local peaks is approximately 0.16 in this example, and the two sets of peaks are offset from each other by approximately 0.008. Presumably the constraint subspace could be defined in this example by these 13 frequencies. However, note that in this example, both dot products are already approximately zero at the frequencies 0.404, 0.436, 0.468, 0.5. Thus the constraint subspace can likely be formed from the remaining nine frequencies 0.412, 0.42, 0.428, 0.444, 0.452, 0.46, 0.476, 0.484, 0.492. At this point it is

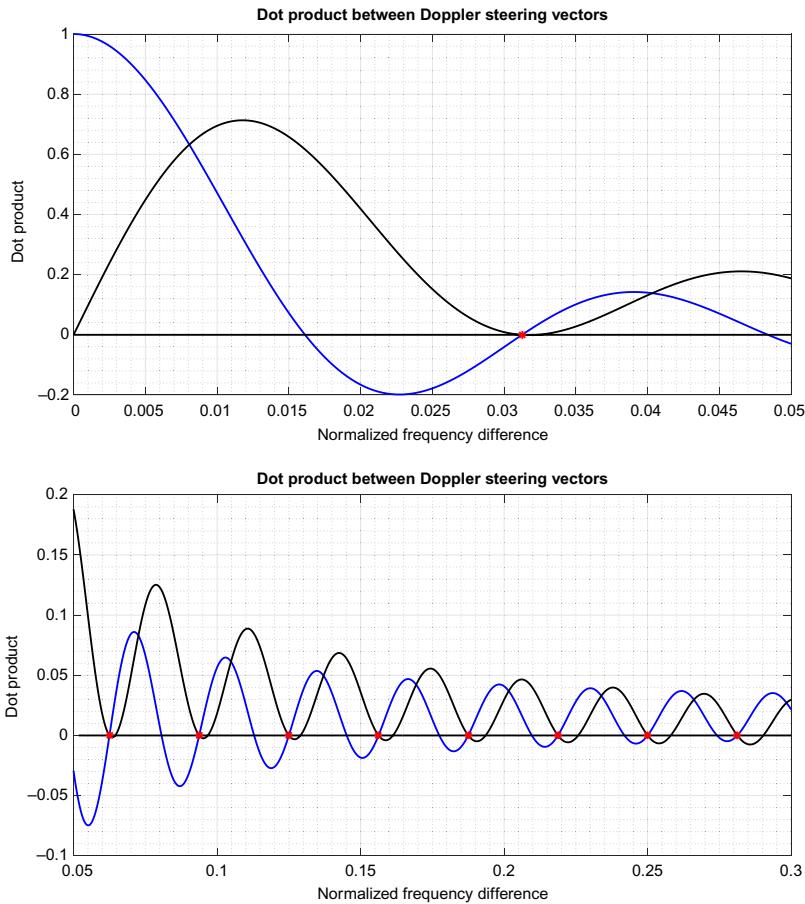
**FIG. 2.35**

Behavior of DFT vectors.

not known if it is necessary to null all of these remaining frequencies to obtain good performance, and more work should be done to explore this problem. Finally, note that this method of identifying the frequencies to define the constraint subspace is not limited to the filter design approach described before. It may also be used in conjunction with the constrained least squares solution to the filter design problem.

#### **2.3.3.4 Generalized Sidelobe Canceller**

Consider a linearly constrained adaptive beamforming problem where the constraints are given by

**FIG. 2.36**

Expanded view of behavior of DFT vectors.

$C^H \mathbf{w}_{\text{complex}} = \mathbf{f}_{\text{complex}}$  (2.538)  
with  $C - N \times k$  complex constraint matrix,  $\mathbf{w}_{\text{complex}} - N \times 1$  complex weight vector, and  $\mathbf{f}_{\text{complex}} - k \times 1$  complex constraint vector.

Let  $\boldsymbol{\varphi}_{\text{complex},1}, \dots, \boldsymbol{\varphi}_{\text{complex},N}$  be the columns of  $C$ :

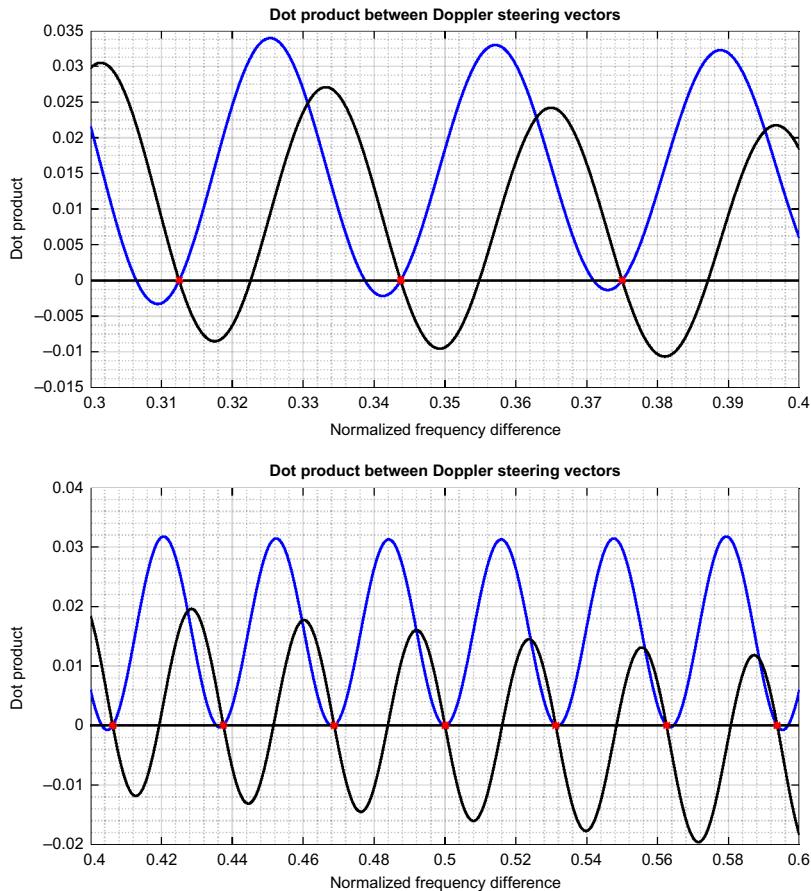
$$C = [\boldsymbol{\varphi}_{\text{complex},1} : \dots : \boldsymbol{\varphi}_{\text{complex},k}] \quad (2.539)$$

Each  $\boldsymbol{\varphi}_{\text{complex},i}, i = 1, \dots, k$  is an  $N \times 1$  vector. The constraints therefore are given by

$$\boldsymbol{\varphi}_{\text{complex},i}^* \cdot \mathbf{w}_{\text{complex}} = f_i, \quad i = 1, \dots, k \quad (2.540)$$

Let  $\boldsymbol{\varphi}_i, \boldsymbol{\varphi}'_i, i = 1, \dots, k$  and  $\mathbf{w}$  be the corresponding  $2N$ -dimensional real vectors. Then in the  $2N$ -dimensional space (3) corresponds to

$$\boldsymbol{\varphi}_i \mathbf{w}_{\varphi_i} = f_i, \quad i = 1, \dots, k \quad (2.541)$$

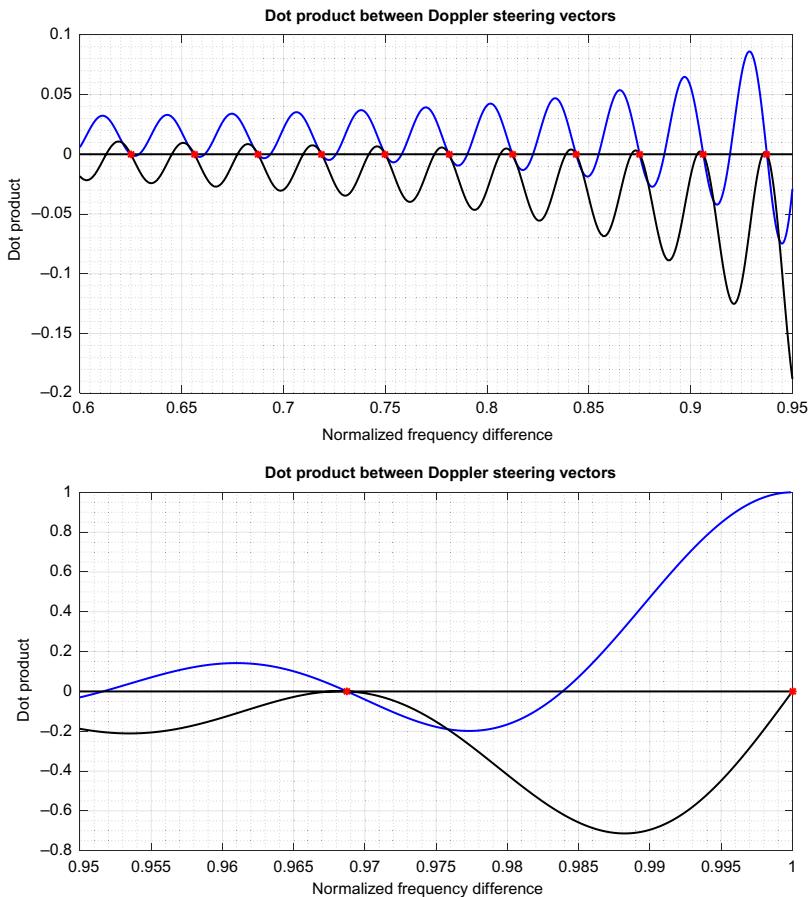
**FIG. 2.37**

Expanded view of behavior of DFT vectors.

where  $\mathbf{w}_{\varphi_i}$  is the projection of  $\mathbf{w}$  into the  $\varphi_i, \varphi'_i$  plane. The geometric product  $\varphi_i \mathbf{w}_{\varphi_i}$  defines the spinor  $f_i$ , which is now seen to operate in the  $\varphi_i, \varphi'_i$  plane. From Eq. (2.541) the projected vector  $\mathbf{w}_{\varphi_i}$  is obtained as

$$\mathbf{w}_{\varphi_i} = \frac{\varphi_i f_i}{|\varphi_i|^2}, \quad i = 1, \dots, k \quad (2.542)$$

Eq. (2.542) gives the projections of  $\mathbf{w}$  into  $k$  linearly independent two-dimensional subspaces of  $R^{2N}$ . Thus the first  $2k$  components of  $\mathbf{w}$  are linear combinations of the constraint vectors  $\varphi_i, \varphi'_i$   $i = 1, \dots, k$  and may be obtained by Gram-Schmidt orthogonalization of the vectors  $\mathbf{w}_{\varphi_i}, i = 1, \dots, k$ . This  $2k$ -component vector will be denoted  $\mathbf{w}_{\parallel}$ .

**FIG. 2.38**

Expanded view of behavior of DFT vectors.

The remaining  $2(N - k)$  components of  $\mathbf{w}$  are free to be specified and form a vector  $\mathbf{w}_{\perp}$  orthogonal to  $\mathbf{w}_{\parallel}$ . Hence write

$$\mathbf{w} = \mathbf{w}_{\parallel} + \mathbf{w}_{\perp} \quad (2.543)$$

The corresponding complex vector may be written

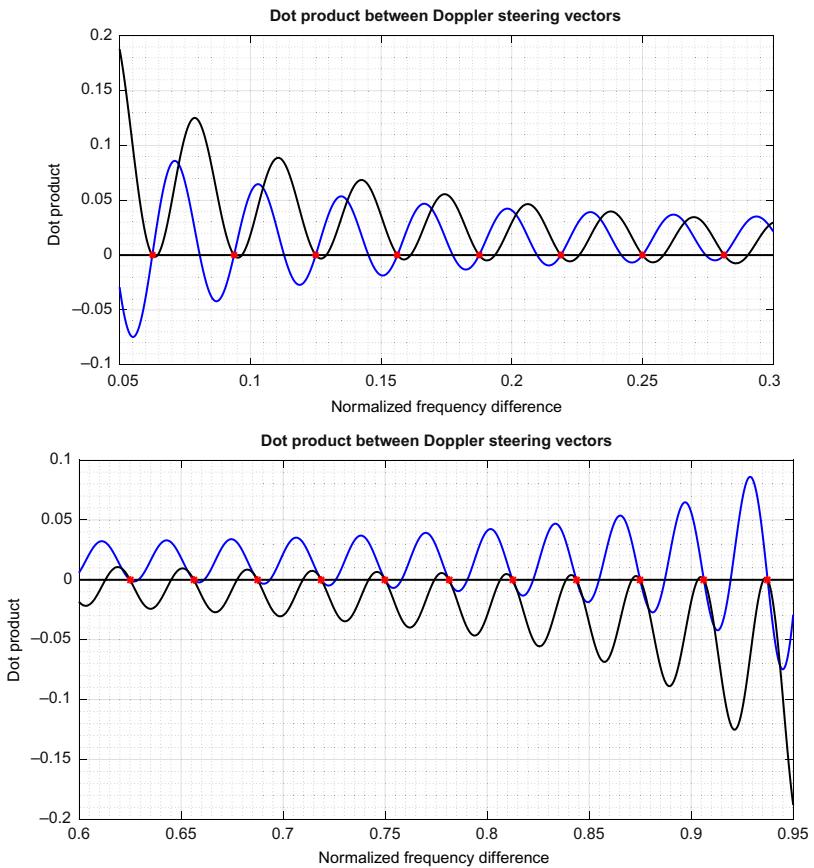
$$\mathbf{w}_{\text{complex}} = \mathbf{w}_{\text{complex}, \parallel} + \mathbf{w}_{\text{complex}, \perp} \quad (2.544)$$

Thus the application of this complex vector to the complex data is given by

$$\mathbf{w}_{\text{complex}}^* \cdot \mathbf{z}_{\text{complex}} = \mathbf{w}_{\text{complex}, \parallel}^* \cdot \mathbf{z}_{\text{complex}} + \mathbf{w}_{\text{complex}, \perp}^* \cdot \mathbf{z}_{\text{complex}} \quad (2.545)$$

In the context of geometric algebra this relation corresponds to

$$\mathbf{w}_{\text{complex}}^* \cdot \mathbf{z}_{\text{complex}} = \mathbf{w}_{\parallel} \mathbf{z}_{\mathbf{w}_{\parallel}} + \mathbf{w}_{\perp} \mathbf{z}_{\mathbf{w}_{\perp}} \quad (2.546)$$

**FIG. 2.39**

Expanded view of behavior of DFT vectors.

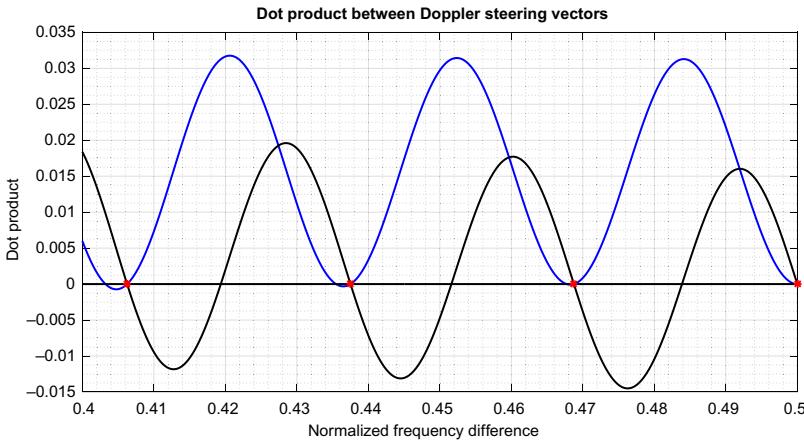
In this expression,  $z_{w_{\parallel}}$  is the orthogonal projection of the data vector  $z$  onto a bivector  $i_{w_{\parallel}}$  whereas  $z_{w_{\perp}}$  is the orthogonal projection of the data vector  $z$  onto a bivector  $i_{w_{\perp}}$ . These bivectors are given by

$$i_{w_{\parallel}} = \frac{\mathbf{w}'_{\parallel} \mathbf{w}_{\parallel}}{|\mathbf{w}'_{\parallel} \mathbf{w}_{\parallel}|} \quad (2.547)$$

$$i_{w_{\perp}} = \frac{\mathbf{w}'_{\perp} \mathbf{w}_{\perp}}{|\mathbf{w}'_{\perp} \mathbf{w}_{\perp}|} \quad (2.548)$$

Observe now that the vector  $\mathbf{w}_{\parallel}$  is a linear combination of the constraint vectors  $\varphi_i, \varphi'_i, i = 1, \dots, k$  and therefore lies in the subspace defined by these constraint vectors. Define this subspace by its unit pseudoscalar:

$$I_{2k} = \frac{\varphi_1 \wedge \varphi'_1 \wedge \dots \wedge \varphi'_k \wedge \varphi'_k}{|\varphi_1 \wedge \varphi'_1 \wedge \dots \wedge \varphi'_k \wedge \varphi'_k|} \quad (2.549)$$

**FIG. 2.40**

Expanded view of behavior of DFT vectors.

Hence  $I_{2k}$  is defined by the columns of the constraint matrix  $C$ . We may write the vectors  $\mathbf{w}_{\parallel}, \mathbf{w}_{\perp}$  as the projection and rejection of  $\mathbf{w}$  from this subspace:

$$\begin{aligned}\mathbf{w} &= (\mathbf{w} \cdot I_{2k}) I_{2k}^\dagger + (\mathbf{w} \wedge I_{2k}) I_{2k}^\dagger \\ &= \mathbf{w}_{\parallel} + \mathbf{w}_{\perp}\end{aligned}\quad (2.550)$$

By this construction it follows that  $\mathbf{w}_{\parallel}$  comprises a linear combination of the constraint vectors  $\boldsymbol{\varphi}_i, \boldsymbol{\varphi}'_i, i = 1, \dots, k$ . Similarly,  $\mathbf{w}_{\parallel}'$  necessarily does as well. Therefore it follows immediately that  $i_{w_{\parallel}}$  is a subspace of  $I_{2k}$ . A similar argument shows that  $i_{w_{\perp}}$  is orthogonal to  $I_{2k}$ . Now also project and reject the data vector  $\mathbf{z}$  to and from this subspace:

$$\begin{aligned}\mathbf{z} &= (\mathbf{z} \cdot I_{2k}) I_{2k}^\dagger + (\mathbf{z} \wedge I_{2k}) I_{2k}^\dagger \\ &= \mathbf{z}_{\parallel} + \mathbf{z}_{\perp}\end{aligned}\quad (2.551)$$

Examine

$$\begin{aligned}\mathbf{z}_{w_{\parallel}} &= (\mathbf{z} \cdot i_{w_{\parallel}}) i_{w_{\parallel}}^\dagger \\ &= (\mathbf{z}_{\parallel} \cdot i_{w_{\parallel}}) i_{w_{\parallel}}^\dagger + (\mathbf{z}_{\perp} \cdot i_{w_{\parallel}}) i_{w_{\parallel}}^\dagger \\ &= \mathbf{z}_{\parallel w_{\parallel}} + (\mathbf{z}_{\perp} \cdot i_{w_{\parallel}}) i_{w_{\parallel}}^\dagger\end{aligned}\quad (2.552)$$

The first term in this expression is the orthogonal projection of  $\mathbf{z}_{\parallel}$  onto the subspace defined by  $i_{w_{\parallel}}$ . However, in the second term  $\mathbf{z}_{\perp}$  is orthogonal to  $I_{2k}$ . Hence it also orthogonal to  $i_{w_{\parallel}}$ . Therefore we find

$$\mathbf{z}_{w_{\parallel}} = \mathbf{z}_{\parallel w_{\parallel}} \quad (2.553)$$

This result says that projecting  $z$  onto  $i_{w_{\parallel}}$  is the same as projecting  $z_{\parallel}$  onto  $i_{w_{\parallel}}$ . By a similar analysis we find that projecting  $z$  onto  $i_{w_{\perp}}$  is the same as projecting  $z_{\perp}$  onto  $i_{w_{\perp}}$ :

$$z_{w_{\perp}} = (z_{\perp} \cdot i_{w_{\perp}}) i_{w_{\perp}}^{\dagger} = z_{\perp w_{\perp}} \quad (2.554)$$

Therefore we find

$$\begin{aligned} w_{\text{complex}}^* \cdot z_{\text{complex}} &= w_{\parallel}^* z_{\parallel w_{\parallel}} + w_{\perp}^* z_{\perp w_{\perp}} \\ &= w_{\text{complex}, \parallel}^* z_{\text{complex}, \parallel} + w_{\text{complex}, \perp}^* z_{\text{complex}, \perp} \end{aligned} \quad (2.555)$$

This result says that the processor  $w_{\text{complex}, \parallel}^* \cdot z_{\text{complex}}$  naturally decomposes into two components—one that operates only on components of  $z$  that are within the constraint subspace and one that operates only on components of  $z$  that are outside of the constraint subspace. The first part of the processor is implemented as the quantity:

$$w_{\text{complex}, \parallel}^* \cdot z_{\text{complex}} = w_{\text{complex}, \parallel}^* z_{\text{complex}, \parallel} \quad (2.556)$$

Note that  $w_{\text{complex}, \parallel}$  has already been specified by the constraints. For any  $z_{\text{complex}} = z_{\text{complex}, \parallel}$  that is *wholly* within the constraint subspace, the constraints are satisfied.

The task of interest now is to choose the vector  $w_{\text{complex}, \perp}$  that implements the second part of the processor:

$$w_{\text{complex}, \perp}^* \cdot z_{\text{complex}} = w_{\text{complex}, \perp}^* z_{\text{complex}, \perp} \quad (2.557)$$

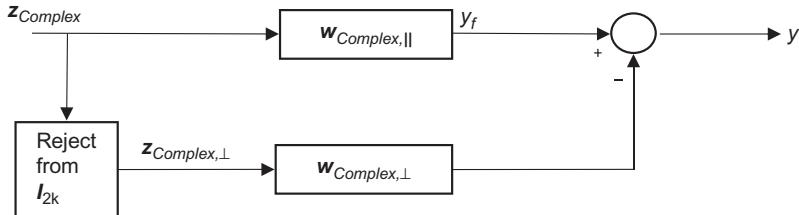
We will select it so that

- Vectors that lie *wholly* within the constraint subspace are processed to insure the constraints are satisfied—this is easy since if a vector is wholly within the constraint space,  $z_{\text{complex}, \perp} = 0$  and there is no response from this part of the processor regardless of how we choose  $w_{\text{complex}, \perp}$ ;
- Any vector that is *not wholly within the constraint subspace*—i.e., if it has any component outside of the constraint subspace—is processed to minimize the total output power from the processor. This requirement determines  $w_{\text{complex}, \perp}$ .

Observe that any vector that is not wholly within the constraint subspace will cause the first component of the processor nonetheless to respond if that vector has a component in the constraint subspace. We will take advantage of the fact that this vector *also* causes the second component of the processor to respond and use this response to minimize the overall response of the processor. Thus the goal is to satisfy the constraints for vectors wholly within the constraint space and to minimize the overall processor response to all other vectors.

Consider an adaptive processor as in Fig. 2.41.

The top leg of this processor is the first component discussed before. It is determined by the constraints and may be thought of as a *fixed beamformer*, i.e., it is not adaptive. Now consider the lower leg. If the input data vector is *wholly within the*

**FIG. 2.41**

Adaptive processor.

constraint subspace  $I_{2k}$ , then as already mentioned  $z_{\text{complex},\perp} = 0$  and the lower leg of the processor does nothing. In this case, the output of the processor is

$$y = \mathbf{w}_{\text{complex},\parallel}^* \cdot \mathbf{z}_{\text{complex}} = \mathbf{w}_{\text{complex},\parallel}^* \cdot \mathbf{z}_{\text{complex},\parallel} \quad (2.558)$$

Observe that because the data vector is wholly within the constraint subspace we have

$$\mathbf{z}_{\text{complex}} = \mathbf{z}_{\text{complex},\parallel} = \sum_{i=1}^k z_i \boldsymbol{\varphi}_{\text{complex},i} \quad (2.559)$$

where  $z_i, i=1, \dots, k$  are the components of  $\mathbf{z}_{\text{complex}}$  along the linearly independent vectors  $\boldsymbol{\varphi}_{\text{complex},i}, i=1, \dots, k$ . Thus

$$\begin{aligned} y &= \mathbf{w}_{\text{complex},\parallel}^* \cdot \mathbf{z}_{\text{complex}} = \sum_{i=1}^k z_i \mathbf{w}_{\text{complex},\parallel}^* \cdot \boldsymbol{\varphi}_{\text{complex},i} \\ &= \sum_{i=1}^k z_i f_i \end{aligned} \quad (2.560)$$

This is the result required by the constraints.

Now consider the case when the input data vector is not wholly within the constraint subspace. In this case

$$\mathbf{z}_{\text{complex}} = \mathbf{z}_{\text{complex},\parallel} + \mathbf{z}_{\text{complex},\perp} \quad (2.561)$$

where  $\mathbf{z}_{\text{complex},\parallel}$  is the component of this input data vector in the constraint subspace and  $\mathbf{z}_{\text{complex},\perp} \neq 0$  is the component of the input data vector orthogonal to the constraint subspace. The response of the upper leg is again

$$y_f = \mathbf{w}_{\text{complex},\parallel}^* \cdot \mathbf{z}_{\text{complex},\parallel} \quad (2.562)$$

Even though this data vector is not wholly in the constraint subspace, the fixed beamformer nonetheless responds to it because the vector has a component in the constraint space. The approach now is to use the lower leg to “cancel” this response. The lower leg response is given by  $\mathbf{w}_{\text{complex},\perp}^* \cdot \mathbf{z}_{\text{complex},\perp}$ . The response of the overall processor is therefore

$$y = \mathbf{w}_{\text{complex},\parallel}^* \cdot \mathbf{z}_{\text{complex},\parallel} - \mathbf{w}_{\text{complex},\perp}^* \cdot \mathbf{z}_{\text{complex},\perp} \quad (2.563)$$

Ideally whenever the input data vector is not wholly within the constraint subspace—i.e., when  $\mathbf{z}_{\text{complex},\perp} \neq 0$ —we would like the response of the processor to be zero:

$$y = y_f - \mathbf{w}_{\text{complex},\perp}^* \cdot \mathbf{z}_{\text{complex},\perp} = 0 \quad (2.564)$$

This would occur if we could choose  $\mathbf{w}_{\text{complex},\perp}$  so that

$$\mathbf{w}_{\text{complex},\parallel}^* \cdot \mathbf{z}_{\text{complex},\parallel} = \mathbf{w}_{\text{complex},\perp}^* \cdot \mathbf{z}_{\text{complex},\perp} \quad (2.565)$$

However, this would require that we choose a different weight vector  $\mathbf{w}_{\text{complex},\perp}$  for each input data vector  $\mathbf{z}$ , and our goal is to choose a *single* vector  $\mathbf{w}_{\text{complex},\perp}$  for processing the data. Hence we cannot do this. Instead, we will choose  $\mathbf{w}_{\text{complex},\perp}$  to minimize the output power of the overall processor:

$$E[|y|^2] = E\left[\left|y_f - \mathbf{w}_{\text{complex},\perp}^* \cdot \mathbf{z}_{\text{complex},\perp}\right|^2\right] \quad (2.566)$$

Note that the constraints are already satisfied in the fixed beamformer of the upper leg of the processor, so we now have *an unconstrained minimization problem*.

To solve this unconstrained minimization problem, examine

$$\begin{aligned} \left|y_f - \mathbf{w}_{\text{complex},\perp}^* \cdot \mathbf{z}_{\text{complex},\perp}\right|^2 &= \left(y_f - \mathbf{w}_{\text{complex},\perp}^H \mathbf{z}_{\text{complex},\perp}\right) \left(y_f^* - \mathbf{z}_{\text{complex},\perp}^H \mathbf{w}_{\text{complex},\perp}\right) \\ &= |y_f|^2 - y_f \mathbf{z}_{\text{complex},\perp}^H \mathbf{w}_{\text{complex},\perp} - \mathbf{w}_{\text{complex},\perp}^H \mathbf{z}_{\text{complex},\perp} y_f^* \\ &\quad + \mathbf{w}_{\text{complex},\perp}^H \mathbf{z}_{\text{complex},\perp} \mathbf{z}_{\text{complex},\perp}^H \mathbf{w}_{\text{complex},\perp} \end{aligned} \quad (2.567)$$

Thus Eq. (2.564) yields

$$E[|y|^2] = E[|y_f|^2] - \mathbf{r}^H \mathbf{w}_{\text{complex},\perp} - \mathbf{w}_{\text{complex},\perp}^H \mathbf{r} + \mathbf{w}_{\text{complex},\perp}^H \mathbf{R}_\perp \mathbf{w}_{\text{complex},\perp} \quad (2.568)$$

where

$$\mathbf{R}_\perp = E\left[\mathbf{z}_{\text{complex},\perp} \mathbf{z}_{\text{complex},\perp}^H\right] \quad (2.569)$$

$$\mathbf{r} = E\left[\mathbf{z}_{\text{complex},\perp} y_f^*\right] \quad (2.570)$$

Taking derivatives in Eq. (2.568) with respect to  $\mathbf{w}_{\text{complex},\perp}$  and setting them equal to zero yields

$$-2\mathbf{r} + 2\mathbf{R}_\perp \mathbf{w}_{\text{complex},\perp} = 0 \quad (2.571)$$

This in turn implies

$$\mathbf{w}_{\text{complex},\perp} = \mathbf{R}_\perp^{-1} \mathbf{r} \quad (2.572)$$

As may be seen,  $\mathbf{R}_\perp$  is simply the covariance matrix of the rejected vector  $\mathbf{z}_{\text{complex},\perp}$  and  $\mathbf{r}$  is the cross-correlation between the rejected vector  $\mathbf{z}_{\text{complex},\perp}$  and the output  $y_f^*$  of the fixed beamformer. To complete the analysis, we must show that  $\mathbf{w}_{\text{complex},\perp}$  is orthogonal to  $\mathbf{w}_{\text{complex},\parallel}$ . We have

$$\mathbf{w}_{\text{complex},\perp}^* \cdot \mathbf{w}_{\text{complex},\parallel} = \mathbf{r}^H \mathbf{R}_\perp^{-H} \mathbf{w}_{\text{complex},\parallel} \quad (2.573)$$

Observe now that  $R_{\perp}$  is a linear transformation acting wholly within the subspace orthogonal to the constraint subspace containing  $\mathbf{w}_{\text{complex},\parallel}$ . It follows therefore that

$$R_{\perp}^{-H}\mathbf{w}_{\text{complex},\parallel} = 0 \quad (2.574)$$

We have obtained an adaptive processor that

- Satisfies the constraints for vectors that are *wholly within* the constraint subspace; and
- Minimizes the output power of the processor for vectors that have *any* component outside the constraint subspace.

The response of the processor as a whole may be written as

$$\begin{aligned} y &= \mathbf{w}_{\text{complex}}^* \cdot \mathbf{z}_{\text{complex}} = \left( \mathbf{w}_{\text{complex},\parallel}^* - \mathbf{w}_{\text{complex},\perp}^* \right) \cdot \mathbf{z}_{\text{complex}} \\ &= \left( \mathbf{w}_{\text{complex},\parallel}^H - \mathbf{r}^H R_{\perp}^{-1} \right) \mathbf{z}_{\text{complex}} \end{aligned} \quad (2.575)$$

Now use

$$y_f^* = \mathbf{z}_{\text{complex}}^H \mathbf{w}_{\text{complex},\parallel} \quad (2.576)$$

This relation leads to

$$\begin{aligned} \mathbf{r} &= E[\mathbf{z}_{\text{complex},\perp} y_f^*] = E[\mathbf{z}_{\text{complex},\perp} \mathbf{z}_{\text{complex}}^H] \mathbf{w}_{\text{complex},\parallel} \\ &= R_{\perp,z} \mathbf{w}_{\text{complex},\parallel} \end{aligned} \quad (2.577)$$

with

$$R_{\perp,z} = E[\mathbf{z}_{\text{complex},\perp} \mathbf{z}_{\text{complex}}^H] \quad (2.578)$$

Thus the overall processor response becomes

$$y = \left( \mathbf{w}_{\text{complex},\parallel}^H - \mathbf{w}_{\text{complex},\parallel}^H R_{\perp,z}^H R_{\perp}^{-1} \right) \mathbf{z}_{\text{complex}} \quad (2.579)$$

Now define

$$R_0^{-1} = I - R_{\perp,z}^H R_{\perp}^{-1} \quad (2.580)$$

The overall processor response becomes finally

$$y = \mathbf{w}_{\text{complex},\parallel}^H R_0^{-1} \mathbf{z}_{\text{complex}} \quad (2.581)$$

This processor is known as the *generalized sidelobe canceller* (GSLC).

To relate this result to the traditional presentation of the GSLC, write

$$\mathbf{z}_{\text{complex},\perp} = \mathbf{B}^H \mathbf{z}_{\text{complex}} \quad (2.582)$$

In this expression,  $\mathbf{B}$  is known as a *blocking matrix*. It blocks the components of  $\mathbf{z}_{\text{complex}}$  that are within the constraint subspace. With this we find

$$\begin{aligned} \mathbf{r} &= E[\mathbf{z}_{\text{complex},\perp} \mathbf{z}_{\text{complex}}^H \mathbf{w}_{\text{complex},\parallel}] \\ &= E[\mathbf{B}^H \mathbf{z}_{\text{complex}} \mathbf{z}_{\text{complex}}^H \mathbf{w}_{\text{complex},\parallel}] \\ &= \mathbf{B}^H \mathbf{R} \mathbf{w}_{\text{complex},\parallel} \end{aligned} \quad (2.583)$$

where  $R$  is the full covariance matrix of  $\mathbf{z}_{\text{complex}}$ . Similarly

$$\begin{aligned} R_{\perp} &= E \left[ \mathbf{z}_{\text{complex}, \perp} \mathbf{z}_{\text{complex}, \perp}^H \right] \\ &= E \left[ B^H \mathbf{z}_{\text{complex}} \mathbf{z}_{\text{complex}}^H B \right] \\ &= B^H R B \end{aligned} \quad (2.584)$$

Thus in this notation

$$\mathbf{w}_{\text{complex}, \perp} = [B^H R B]^{-1} B^H R \mathbf{w}_{\text{complex}, \parallel} \quad (2.585)$$

The constraints can be satisfied by computing  $\mathbf{w}_{\text{complex}, \parallel}$  as

$$\mathbf{w}_{\text{complex}, \parallel} = C [C^H C]^{-1} \mathbf{f}_{\text{complex}} \quad (2.586)$$

Eqs. (2.585) and (2.586) give the traditional formulation of the generalized sidelobe canceller as discussed, for example, in Ref. [56].

From a geometric perspective, the GSLC operates as follows:

1. The incoming data is processed with a fixed beamformer that enforces the constraints;
2. The incoming data is rejected from the subspace defined by the constraints and processed by determining the covariance matrix of the rejected data as well as its cross-correlation with the output of the fixed beamformer;
3. The processed rejected data is subtracted from the output of the fixed beamformer.

## 2.4 CONCLUSION—FUTURE RESEARCH OPPORTUNITIES

This work introduces ideas from geometric algebra and applies them to some basic radar signal processing problems. The primary goal has been to introduce (or perhaps better to reintroduce) geometric ways of thinking into the solution of signal processing problems involving complex data. Although geometric algebra has been applied extensively to physical problems both in three-dimensional space and in four-dimensional Minkowski spacetime, its application to general  $N$ -dimensional complex vector spaces or their corresponding  $2N$ -dimensional real vector spaces has been pursued much less by comparison, despite the fact that the techniques are generally applicable in any dimension. This lack of application by the signal processing community is almost certainly due to a lack of exposure to the underlying ideas of geometric algebra. It is the author’s hope that the present work will expose a larger group of engineers to these ideas and thereby initiate new lines of thinking in approaching signal processing challenges.

With respect to the topics discussed herein, areas of potential further exploration include the following:

- Further studies of the DFT may lead to analogous geometric interpretations of related results such as the fractional discrete Fourier transform, which is in turn is related to linear frequency-modulated chirp signal processing.
- The geometric approach to detection processing is ripe for further exploration. Other detectors based on this approach have been discussed in Ref. [23]. Performance assessments, implementation issues, and adaptive approaches based on geometric thinking are all areas that merit further exploration.
- Characterization of the relationship between the geometric orthogonality principle and least-squares processing in the context of geometric algebra may lead to new perspectives on this very old aspect of statistical signal processing.
- Reformulating conjugate direction and multistage Wiener filters and related approaches in the language of geometric algebra will likely give rise to heretofore unforeseen insights into filtering problems. See Goldstein et al. [47] and Scharf et al. [55] for geometric formulations of these important classes of filters.
- In 2004 Daum suggested that Lie algebras may lead to advances in tracking and filtering that go beyond the Kalman filter [44]. Lie groups and algebras can be formulated in terms of geometric algebra [18], so it may be fruitful to recast tracking problems in terms of geometric algebra via Lie groups/algebra to seek geometric extensions in this very important area of radar signal processing.

The previous listing provides just a small indication of the areas where geometric thinking may lead to new insights and approaches. More generally, most areas of radar signal processing involving complex data have been attacked from an algebraic perspective with vectors and matrices being the fundamental tools of application. Recasting these problems in the language of geometric algebra using multivectors and spinors as the basic tools of interest will undoubtedly lead to new geometrical ways of thinking about these problems and may lead to useful new implementations and solutions.

## REFERENCES

Geometric algebra is a large body of mathematics with an extensive body of literature. The following listing provides only a small sampling of the growing literature in this area that may be of interest to the radar engineer.

### Books

- [1] J.W. Arthur, *Understanding Geometric Algebra for Electromagnetic Theory*, IEEE Press, Piscataway, New Jersey, 2011.
- [2] W.E. Baylis, *Electrodynamics: A Modern Geometrical Approach*, Springer-Verlag New York, Inc., New York, New York, 1999.
- [3] E. Bayro-Corrochano, G. Sobczyk (Eds.), *Geometric Algebra with Applications in Science and Engineering*, Springer-Verlag New York, Inc., New York, New York, 2001.
- [4] E. Bayro-Corrochano, G. Scheuermann (Eds.), *Geometric Algebra Computing in Engineering and Computer Science*, Springer-Verlag, London, 2010.

- [5] C. Doran, A. Lasenby, *Geometric Algebra for Physicists*, Cambridge University Press, United Kingdom, 2003.
- [6] L. Dorst, C. Doran, J. Lasenby (Eds.), *Applications of Geometric Algebra in Computer Science and Engineering*, Springer Science and Business Media, New York, 2002.
- [7] D. Hestenes, *Space-Time Algebra*, Gordon and Breach, New York, 1966.
- [8] D. Hestenes, G. Sobczyk, *Clifford Algebra to Geometric Calculus*, D. Reidel Publishing Company, Dordrecht, 1984.
- [9] D. Hestenes, *New Foundations for Classical Mechanics*, second ed., Kluwer Academic Publishers, Dordrecht, 1999.
- [10] P. Lounesto, *Clifford Algebras and Spinors*, Cambridge University Press, United Kingdom, 2001.
- [11] A. MacDonald, *Linear and Geometric Algebra*, CreateSpace Independent Publishing Platform, Charleston, SC, 2010.
- [12] A. MacDonald, *Vector and Geometric Calculus*, CreateSpace Independent Publishing Platform, Charleston, SC, 2012.
- [13] C. Perwass, *Geometric Algebra with Applications in Engineering*, Springer, 2009.
- [14] G. Sobczyk, *New Foundations in Mathematics: The Geometric Concept of Number*, Birkhäuser Mathematics, 2013.

### Introductory and/or General Papers

- [15] E.F. Bolinder, *Clifford Algebra: What Is It?* IEEE Antennas Propag. Soc. Newslett. (August 1987).
- [16] M. Buchanan, *Geometric intuition*, Nat. Phys. 7 (2011) (Corrected online 28 October 2011), 442.
- [17] J.M. Chappell, S.P. Drake, C.L. Seidel, L.J. Gunn, A. Iqbal, A. Allison, D. Abbott, *Geometric algebra for electrical and electronic engineers*, Proc. IEEE 102 (9) (2014) 1340–1363.
- [18] C. Doran, D. Hestenes, F. Sommen, N. Van Acker, *Lie groups as spin groups*, J. Math. Phys. 34 (8) (1993) 3642–3669.
- [19] S. Gull, A. Lasenby, C. Doran, *Imaginary numbers are not real – the geometric algebra of spacetime*, Found. Phys. 23 (9) (1993) 1175–1201.
- [20] D. Hestenes, *A unified language for mathematics and physics*, in: J.S.R. Chisolm, A. K. Commons (Eds.), *Clifford Algebras and Their Applications in Mathematical Physics*, D. Reidel Publishing Company, 1986.
- [21] D. Hestenes, *Oersted medal lecture 2002: reforming the mathematical language of physics*, Am. J. Phys. 71 (2) (2003) 104–121.
- [22] L.V. Meisel, *A Mathematical Formulation of Geometric Algebra in 3-Space*, Technical Report ARCCB-TR-95016, US Army Armament Research, Development and Engineering Center, 1995.
- [23] K.J. Sangston, *Geometry of complex data*, IEEE AES Syst. Mag. Tutor. 31 (3) (2016) 32–69.
- [24] I. Stewart, *Hermann Grassmann was right*, Nature 321 (1986) 17.
- [25] T.G. Vold, *An introduction to geometric algebra with an application in rigid body mechanics*, Am. J. Phys. 61 (6) (1993) 491–504.
- [26] T.G. Vold, *An introduction to geometric calculus and its application to electrodynamics*, Am. J. Phys. 61 (6) (1993) 505–513.

## Signals, Transforms, and Signal Processing Applications

- [27] D. Alfsmann, H. Gockler, S. Sangwine, T. Ell, Hypercomplex algebras in digital signal processing: benefits and drawbacks, 15th European Signal Processing Conference, Poland, September 3–7, 2007.
- [28] T. Batard, M. Berthier, Spinor Fourier transform for image processing, *IEEE J. Sel. Top. Signal Process.* 7 (4) (2013) 605–613.
- [29] E. Bayro-Corrochano, Y. Zhang, The motor extended Kalman filter: a geometric approach for rigid motion estimation, *J. Math. Imaging Vision* 13 (2000) 205–228.
- [30] R.S. Bucy, Signal space geometry, *Inform. Sci.* 40 (1986) 75–82.
- [31] R.S. Bucy, Geometry and multiple direction estimation, *Inform. Sci.* 57–58 (1991) 145–158.
- [32] T. Bulow, G. Sommer, Hypercomplex signals – a novel extension of the analytic signal to the multidimensional case, *IEEE Trans. Signal Process.* 49 (11) (2002) 2844–2852.
- [33] M. Felsberg, G. Sommer, The multidimensional isotropic generalization of quadrature filters in geometric algebra, in: Algebraic Frames for the Perception-Action Cycle, vol. 1888, Lecture Notes in Computer Science, 2000, pp. 175–185.
- [34] M. Felsberg, G. Sommer, The monogenic signal, *IEEE Trans. Signal Process.* 49 (12) (2001) 3136–3144.
- [35] S. Hinidama, U. Gamage, J. Lasenby, Optimal estimation and tracking of general rotations using geometric algebra with applications in computer vision, *Proceedings of SPIE, Vision Geometry IX*, vol. 4117, 2000, pp. 261–271.
- [36] M. Ilchenko, T. Narytnik, R. Didkovsky, Clifford algebra in multiple-access noise-signal communication systems, *Telecommun. Radio Eng.* 72 (18) (2013) 1651–1663.
- [37] Y. Kuroe, T. Nitta, E. Hitzer, Applications of Clifford's Geometric Algebra, *SICE J. Control Meas. Syst. Integr.* 4 (1) (2011) 1–10.
- [38] A. Rockwood, D. Hildebrand, Engineering graphics in geometric algebra, in: E. Bayro-Corrochano, G. Scheuermann (Eds.), *Geometric Algebra Computing*, Springer, 2010.
- [39] R.O. Schmidt, New mathematical tools in direction finding and spectral analysis, *Proc. SPIE 27th Ann. Symp.*, August 23, 1983, pp. 7–19.
- [40] R.O. Schmidt, Multilinear array manifold interpolation, *IEEE Trans. Signal Process.* 40 (4) (1992) 857–866.
- [41] J. Stanway, J. Kinsey, Sensor alignment using rotors in geometric algebra, *IEEE International Conference on Robotics and Automation*, Shanghai, China, May 9–13, 2011, pp. 994–999.
- [42] D. Wu, Z. Wang, Strapdown INS/GPS integrated navigation using geometric algebra, *Adv. Appl. Clifford Algebr.* 23 (2013) 767–785.

## Other References

- [43] E. Conte, M. Longo, G. Ricci, Asymptotically Optimum Radar detection in compound-gaussian clutter, *IEEE Trans. Aerosp. Electron. Syst.* 31 (2) (1995) 617–625.
- [44] F. Daum, Nonlinear filters: beyond the Kalman filter, *IEEE AES Syst. Mag. Tutor.* 20 (8) (2005) 57–69.
- [45] A.B. Gershman, G.V. Serebryakov, J.F. Bohme, Constrained Hung-Turner adaptive beamforming algorithm with additional robustness to wide-band and moving jammers, *IEEE Trans. Antennas Propag.* 44 (3) (1996) 361–367.

- [46] F. Gini, Sub-optimum coherent radar detection in a mixture of K-distributed and Gaussian clutter, IEE Proc. Radar Sonar Navig. 144 (1) (1997) 39–48.
- [47] J.S. Goldstein, I.S. Reed, L.L. Scharf, A multistage representation of the wiener filter based on orthogonal projections, IEEE Trans. Inf. Theory 44 (7) (1998) 2943–2959.
- [48] E.K.L. Hung, R.M. Turner, A. Fast Beamforming, Algorithm for large arrays, IEEE Trans. Aerosp. Electron. Syst. 19 (4) (1983) 598–607.
- [49] V.A. Korado, Optimum detection of signals with random parameters against the background noise of unknown intensity under conditions of constant false alarm probability, Radio Eng. Electron. Phys. 13 (6) (1968) 969–972.
- [50] W.L. Melvin, A STAP overview, IEEE AES Syst. Mag. 19 (1) (2004) 19–35 (Special Tutorials Issue).
- [51] B. Picinbono, G. Vezzosi, Detection d'un signal certain dans un bruit non stationnaire et Gaussien, Ann. Telecommun. 25 (1970) 433–439.
- [52] H.V. Poor, An Introduction to Signal Detection and Estimation, Springer-Verlag, 1988.
- [53] L. Scharf, D. Lytle, Signal detection in gaussian noise of unknown level: an invariance application, IEEE Trans. Inf. Theory 17 (4) (1971) 404–411.
- [54] L. Scharf, Statistical Signal Processing, Addison-Wesley, 1991.
- [55] L.L. Scharf, E.K.P. Chong, M.D. Zoltkowski, J.S. Goldstein, I.S. Reed, Subspace expansion and the equivalence of conjugate direction and multistage Wiener filters, IEEE Trans. Signal Process. 56 (10) (2008) 5013–5019.
- [56] B.D. Van Veen, K.M. Buckley, Beamforming: a versatile approach to spatial filtering, IEEE ASSP Mag. (April 1988) 4–24.
- [57] M. Zatman, Comment on ‘Constrained Hung-Turner adaptive beamforming algorithm with additional robustness to wide-band and moving jammers’, IEEE Trans. Antennas Propag. 46 (12) (1998) 1897–1898.

# Foundations of cognitive radar for next-generation radar systems

# 3

**Nathan A. Goodman**

*The University of Oklahoma, Norman, OK, United States*

---

## 3.1 BACKGROUND

Radar technology has improved over many decades. Improvements in transmitter and amplifier technology, precision mixers and oscillators, wideband components, adaptive signal processing, high-throughput digital hardware, digital processing hardware, and other technologies have all dramatically impacted radar capability, usefulness, and reliability. But while these technologies have increased the fidelity and resolution of radar measurements, they have not necessarily impacted the fundamental manner in which radar systems collect their data. By this statement, we mean that most fielded radar systems still capture data according to rigid, predefined patterns using a limited selection of waveform types and parameters. Of course, there are strong historical and technical reasons for this rigidity, including the inherent difficulty in reconfiguring custom processing modules, the need to carefully control data rates between subsystems, and the real-time processing that would be required to implement a radar scheduler that does more than select from among a small collection of modes. As a consequence, radar performance improvement has been driven by advances in component quality and capability, and by the processing algorithms that exploit the radar data. Reduced harmonics, better calibration, increased instantaneous bandwidth, enhanced oscillator stability, and increased dynamic range all lead to processed results that more faithfully represent and characterize the radar system's external scattering and interference environment.

Yet despite all these advances, radar systems still typically operate in the same rigid manner as they have for decades. Wide-area searches are performed in fixed search patterns, track updates are scheduled at standard intervals regardless of current conditions (e.g., the likelihood of dropped or mixed tracks due to the density of potential targets), and multifunction aperture sharing is rudimentary or nonexistent. Advanced adaptive processing algorithms, powered by extraordinary advances in computing capability, are better than ever before at estimating and rejecting interference in order to accomplish desired detection, parameter estimation, and imaging tasks. Any exploitation algorithm, however, can only be as good as the data that it has

to work with, and few systems proactively optimize the quality of the data received—at least not to the extent envisioned by the research community working to develop the new paradigm of cognitive radar.

Consider, for example, an individual driving a car down a street during the daytime. There are many stimuli, cues, constraints, and perils that must be absorbed, verified, and evaluated in order to operate the vehicle safely and productively. The driver primarily uses hearing and vision to assess the environment while adjusting the vehicle's position, speed, and direction. When approaching an intersection, the driver will pay additional attention to the traffic control lights and to other traffic at the intersection. The traffic control light might change at any time; hence, the update rate on the driver's visual observation of the control light will increase so that the time delay between a control light change and observing the change can be minimized. The driver might even slow down (or speed up!) to ensure that a sudden traffic control change will result in a safe stop (or in passing through the intersection). Moreover, the driver will change his or her sensitivity to a potential traffic control change depending on other recent observations and situational awareness. If the driver has observed the control light from far off and knows that the light has been green for an extended period of time, then a control light change may be imminent and the driver will give extra attention. On the other hand, if the light has only recently changed, then another immediate change is not expected, and the driver can focus on other factors. Another example would be if the driver observed children playing in a yard to the side of the road. As the driver approaches, he or she might provide extra attention to the children in case one of them chases a ball into the street. The driver might even take the proactive control step of moving into another lane to increase the available reaction time.

Although the previous examples are not related to radar, they are examples of adaptive sensing. The driver is consistently sensing the environment, but more than that, the driver is adaptively and intelligently using the available sensing resources to optimize information gained, minimize risk, and enhance the chances for good outcomes. The sensing timeline is finite, and any time spent viewing or listening to important features of the external driving environment necessarily means that the same time cannot be focused on other important features. Moreover, the environmental sensing is adaptive in the sense that the driver's behavior strongly depends on other recent observations. And because these observations depend on a dynamic environment, the sensing behavior cannot be fully planned in advance.

With the previous background in mind, cognitive radar is presented as a new paradigm for radar operation. A typical radar system is constantly observing its environment, inferring properties about the environment based on its observations, and modifying its observation of the environment through changing waveforms and antenna pointing direction. In contrast to the human sensing examples, however, few radar systems currently exploit recent observations to form real-time hypotheses about decisions and/or events of interest and then turn those hypotheses into adjustments of the sensing procedure. The ability to feed measurements back into a system computer that will use these measurements and changing external conditions to

optimize future measurements is a unique feature that helps to distinguish cognitive radar. In other words, cognitive radar should have an aspect of adaptive transmission or measurement, which can be interpreted in a number of ways including adaptive selection of waveform parameters, real-time design of the waveform's temporal and spectral structure, adaptive use of transmit beamsteering and/or polarization, reconfigurable apertures on transmit, adaptive scheduling, and even platform maneuvers to enhance performance or survivability.

In Ref. [1], this continual process of learning about the environment through observation, inferring properties of the environment, and modifying sensing parameters was called the perception-action cycle. The radar system constantly perceives its environment, updating its understanding of potential targets, priorities, and sources of interference. The system then takes this updated understanding and turns it into action through waveform optimization, adaptive aperture scheduling, flight profile changes, and/or other actions that change the nature of subsequent received data. This chapter primarily treats the perception step via a Bayesian structure where probability distributions on targets and interference are updated in a sequential nature in response to recent measurements. The action cycle is then an optimization of parameters such as waveform spectrum and beam direction in order to maximize metrics such as mutual information (MI) or signal-to-noise ratio (SNR).

Much of this chapter is focused on the *action* part of the cycle and the inherent potential benefits of closed-loop sensing. For example, the chapter includes optimization strategies for waveforms and transmit beamforming based on statistical representations of the operational environment. However, the *perception* part of the cycle is a necessary part of the feedback loop as well. In this context, perception denotes the ability of the radar system to collect the data it receives, integrate it with other databases or available information sources, and learn about the radar environment. This learning is manifested as an updated understanding of the radar environment's status quo, which can then be acted on to improve future sensing performance. Therefore the perception step can involve spectrum sensing [2,3], calculating echoic flow for navigation [4], updating target probabilities [5,6] or target tracks [7], estimating target maneuverability, or any other method to extract understanding of the scenario.

Cognitive radar is inherently difficult to define, but can be viewed as an intelligent sensing system that continually interrogates the environment in order to achieve its objectives. Cognitive radar uses prior information, measurements from other sensors, and its own collected data to learn about the environment and make real-time decisions about how to proceed. Waveforms can be modified in response to priorities and previous data, and platform flight profiles can be modified to complement other sensors or to anticipate future actions taken by important targets. Multiple sensor platforms can dynamically organize into jointly operating systems and then disband again when collaborative operation is no longer useful.

An early paper on the topic [1] stated that a cognitive radar "...transmitter adjusts its illumination of the environment in an intelligent manner." Moreover, Haykin [1]

describes that a cognitive radar is unique in three important ways: the receiver continuously learns about the environment, the transmitter adjusts its illumination to account for “practical matters” that might be gleaned through the receiver’s previous interactions with the environment, and the system is dynamic and closed with a feedback loop from the receiver to the transmitter. These early thoughts regarding cognitive radar clearly focus on the availability of an adaptive transmitter that learns from prior measurements and adjusts subsequent illuminations accordingly. The timescale of these earlier observations can vary significantly, from measurements made in prior years to measurements as recent as a coherent processing interval (CPI) collected just a few fractions of a second earlier. Indeed, a human-in-the-loop form of cognition might involve analysis of data sets collected over a number of years in order to design new waveforms that give better performance in subsequent data campaigns. In this chapter, we focus on short timescales (on the order of seconds) where an onboard processor must be programmed to serve as the real-time or near real-time cognitive engine. The ability to characterize, preserve [8], update, and quantify information for future use is an inherent feature of cognitive radar, but the practical implementation of this feature in real applications still requires major advances and development.

---

## 3.2 EARLY RESEARCH CONTRIBUTIONS

Some of the earliest work relevant to cognitive radar investigated the notion of adaptive energy allocation to a parallel set of detection decisions [5,9]. In this problem setup, a set of  $N$  detection decisions must be made, and each decision is modeled as a statistically independent channel. It is assumed that the observing sensor has independent control over the illumination of each channel, subject to an overall energy constraint on the illumination. As more illumination energy is allocated to a particular channel, SNR improves and the detection decision becomes more reliable. However, the finite energy constraint requires that energy should be intelligently allocated, which is achieved in Ref. [5] via an optimization of MI at each stage of a sequential set of illuminations. The results in Ref. [5] showed faster convergence to detection decisions, which implies more efficient use of available transmit power or radar timeline. We provide more analysis of adaptive sensing for canonical detection problems later in this chapter.

Building on the philosophies presented in Refs. [1,5,8,9], the work in Ref. [6] combined optimum waveform design [10–13] and sequential hypothesis testing [14–20] to investigate adaptive sensing for target recognition. Posed as a system identification problem, the authors applied previous work in information-based [10] and SNR-optimal waveform design [11,12] to enhance the separation of different target impulse responses given constraints on waveform energy, bandwidth, and time. The target was sequentially illuminated by an optimized waveform at each step, and a sequential probability ratio test (SPRT) [16] was used to determine when to end

the illuminations. The adaptive waveform strategy resulted in faster convergence to accurate decisions compared to traditional waveforms designed to have “good” range autocorrelation responses (i.e., high resolution and low range sidelobes). The optimized waveforms, designed at each step using MI or SNR metrics, often comprise a single narrowband tone or a set of just a few noncontiguous narrowband tones, such that the range autocorrelation responses of the optimized waveforms would not yield pleasing range profiles via matched filtering. On the other hand, the waveforms performed very well for the specific identification task presented to the radar system.

Inspiration for adaptive waveform modulation during radar operation has also been found by observing the echolocation behavior of bats and other mammals [21–23]. Observations of echolocating mammals while navigating or hunting reveals sophisticated control of their waveform’s modulation (analyzed via spectrogram), power, and repetition rate [23]. Moreover, bats appear to optimally adapt their flight trajectories to maximize hunting success. Analysis of bat echolocation waveforms and flight trajectories using traditional radar analysis methods has demonstrated that bats implement real-time adaptive sensing that should be the envy of today’s rigidly operating radar systems. In the time since cognitive radar concepts were proposed in the mid-2000s, there have been a number of published papers analyzing bat echolocation and trying to mimic it in radar systems [24–26].

There has also been significant research into knowledge-aided signal processing [27], which certainly has applications to cognitive radar, but seems to be more restrictive. Knowledge-aided sensing and signal processing implies that the data have already been collected and that additional knowledge sources are being used to better inform the processing algorithms. For example, external databases can be used to identify discrete clutter reflections [28] or to identify range bins that should be excluded from space-time adaptive filter training [29]. Certainly these examples represent the injection of knowledge into a radar and contribute to the concept of a cognitive radar. However, knowledge-aided processing typically does not imply a feedback loop to the transmitter in order to modify future transmissions. This chapter focuses more on the ideas related to modifying the actual measurements received via changes in waveform shaping, waveform parameters, transmit beampatterns, and any other degree of freedom available on transmit. Therefore a feedback path from processor to transmitter is essential.

---

### 3.3 ENABLING HARDWARE AND PROCESSING TECHNOLOGIES

Cognitive sensing has received much more attention in the past decade. This additional attention is due, in part, to advances in hardware technologies that enable adaptive transmitters and to computing technologies that can support the

computation needed to implement algorithms for cognition. For example, the prospects for adaptive waveform modulation have improved with advances in high-fidelity arbitrary waveform synthesis, and adaptive frequency hopping (e.g., to find uncrowded spectrum) is enabled by new software-defined architectures designed to allow frequency agility. Older architectures or systems having mechanically scanned antennas, analog beamforming networks, fixed intermediate and radio frequencies, and analog-based waveform generation do not provide the flexibility necessary to implement an adaptive transmitter that responds to feedback from a high-performance receiver and signal processor. To achieve the full benefit of cognitive radar techniques, the transmitter must be able to respond to recent measurements by rapidly designing a custom waveform and accurately synthesizing it for the next set of measurements. Likewise, adaptive beamsteering techniques or adaptive multibeam concepts require fast electronic access to subarrays (or even individual elements) within an overall aperture. Thus as apertures become more digital, as direct digital-to-analog synthesis of waveforms becomes more commonplace, and as architectures for software-defined systems show that they can respond and retune ever more quickly, there is tremendous growth in the potential for cognitive radar techniques to take advantage of these new degrees of freedom.

All-digital apertures may represent the pinnacle of hardware technology for agile sensing. All-digital apertures comprise an array of elements, with each element having its own independently controlled waveform on transmit and its own analog-to-digital converter (ADC) on receive. Of particular interest to cognitive radar with adaptive transmission, all-digital radar allows an independently defined waveform on every element of the system. In theory, different modulations can be applied to different elements, representing the ultimate in space-time-frequency sensing agility. An all-digital aperture's timeline can be optimally divided into opportunities for full-aperture high-gain single beams, for multiple-input multiple-output (MIMO) operation, for segmented apertures that enable multifunction operation or frequency, polarization, and modulation diversity, for novel space-time radar waveform coding, and so on. The possibilities are endless, but require foundational theory and algorithms for the control and full exploitation of the available degrees of freedom.

Finally, back-end computing for cognitive radar must move away from pipelined and hardwired processing. For example, if a cognitive radar system allows for adaptive segmentation of an aperture, then any hardwired processing that automatically forms beams from the entire aperture may be inappropriate. Similarly, processors for a cognitive radar system may need to support an arbitrary number of radar pulses in a CPI, and if the modulation is arbitrary, then processors will not be able to store fixed, calibrated filter coefficients in memory. Indeed, calibration of adaptive transmitters is an important issue to be considered during implementation, as in some cases the performance improvements achieved through cognitive radar may be very sensitive to small errors in the optimized illumination.

---

## 3.4 SIGNAL PROCESSING FOUNDATIONS FOR COGNITIVE RADAR

As described thus far, cognitive radar involves characterizing and quantifying a radar system's external environment, feeding that information to an adaptive transmitter, and exploiting the information to improve the information contained in subsequent measurements. Obviously, the cognitive radar system requires some way to represent the knowledge that it has obtained, to update the knowledge when additional measurements or other sources of information are obtained, and to then turn that information into an improved future measurement. But as soon as we say that the radar system will represent information, we must define what the information is about. Is the information about potential targets that need to be detected or ruled out? Is the information about various target locations that need to be updated, or about the likelihood that a target belongs to a particular class or type of target? Is there information regarding the potential terrain where the target and/or radar is located, and is there a potential for shadowing? Is the information about an external interference such as clutter or jamming?

It quickly becomes evident that as soon as we start talking about cognition, knowledge, and information, we must define what we are learning about. And once we define what we are learning about, we inherently begin discussing specific radar exploitation tasks such as detection, tracking, and classification. The methods used to represent radar cognition and to optimize or modify radar behavior will, therefore, be highly application dependent and scenario dependent. Different representation schemes and optimization metrics will be appropriate for different applications, and it is difficult to describe a single unifying theory of cognitive radar. Many of the techniques that have been studied in the literature are fundamentally Bayesian in nature, which admits the concepts of prior knowledge as represented through pdfs and other statistical measures, knowledge updates through implementations or approximations to Bayes' Theorem [30], nuisance parameters and interference incorporated into the pdfs of the cognitive radar's sensing model, and MI related to a particular exploitation task. Bayesian methods are not the only alternatives, however, so we will try to include non-Bayesian methods where appropriate in this chapter.

In the following sections, we describe a few of the foundational signal processing concepts that have been used to explore cognitive radar, including sequential hypothesis testing, task-specific MI, and optimized or matched-illumination waveform design.

### 3.4.1 WAVEFORM DESIGN

A critical component of cognitive radar is optimized waveform design. Optimized waveform design can enable improved target recognition, enhance SNR by illuminating target resonances and avoiding frequency bands with strong interference,

reduce waveform cross-correlation for MIMO or dual polarization modes, place transmit nulls on discrete clutter, and can be a vital diversity technique in electronic warfare applications. In theory, frequency-domain methods for waveform optimization can also be used to optimize spatial illumination patterns, although the required array technology is not as mature as the synthesis technologies used to produce arbitrary temporal modulation. In this section, we review important results related to waveform optimization via MI and SNR objective metrics. Optimization of SNR is considered in detail in Ref. [31] using a discrete-time model, so here we are brief and emphasize frequency-domain interpretations.

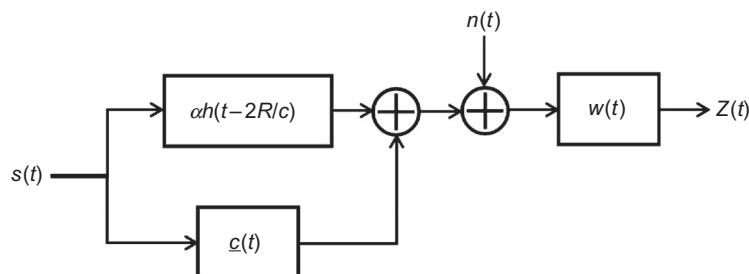
The typical problem formulation for optimum waveform design models the target response as a linear, time-invariant system with impulse response  $h(t)$ . Depending on the specific application, prior knowledge available, and other factors, the impulse response can be modeled as either a deterministic or a random quantity. In Ref. [10], the impulse response was taken as a known deterministic quantity for the SNR-optimization problem, and as a Gaussian random process for optimization of MI. In Ref. [32], both target types were considered under various situations. The usual model considers the received signal to be the result of waveform interaction with the target response, waveform interaction with other undesired external objects such as clutter, plus interference that does not depend on the waveform modulation.

### 3.4.1.1 Deterministic, Known Target Impulse Response

Let  $h(t)$  be a known target impulse response that models an extended radar target as a linear, time-invariant system (see Fig. 3.1). We assume that the operating frequency range of interest is being represented at baseband; hence,  $h(t)$  and all other quantities are assumed to be complex valued. For a target located at a distance  $R$  meters away from a monostatic radar system, the portion of the received signal due to the target is

$$y_h(t) = \alpha s(t) * h(t - 2R/c)$$

where the symbol “ $*$ ” denotes the convolution operation and the constant  $\alpha$  accounts for the strength of the received signal due to radar range equation parameters such as



**FIG. 3.1**

Linear system model for radar measurements comprising target reflection, clutter contributions, additive noise, and receive filtering.

spreading loss and antenna gain. In a similar way, the portion of the received signal that depends on the transmit waveform interacting with external interfering objects (so-called signal-dependent interference) is

$$y_c(t) = s(t) * \underline{c}(t)$$

where the impulse response of the external signal-dependent interference is modeled as a stationary, zero-mean, circularly symmetric Gaussian random process with power spectral density (PSD) denoted as  $\Psi_{cc}(F)$ . In reality, the external clutter is not statistically stationary; for example, the strength of the clutter must decrease with range due to propagation losses, and the structure of the clutter PSD will vary due to changes in the scattering surface. Here, we assume that the clutter is stationary over the region where the target is located. The target response, clutter, and additive interference are passed through an analog receive filter  $w(t)$  before being sampled at the output of the filter. Of course, it is also possible to have an implementation where the signal is sampled by an ADC prior to application of a digital filter, but for the current derivation, the analog filter model is simpler. The output of the receive filter is then

$$z(t) = w(t) * [s(t) * (\alpha h(t - 2R/c) + \underline{c}(t)) + n(t)]. \quad (3.1)$$

The goal of optimized waveform design via an SNR metric is to maximize the SNR contained in  $z(t)$  at a particular sample time  $t = t_M$ . Separating the filter output  $z(t)$  into signal and interference components, the ratio of signal power to interfering power at  $t = t_M$  is

$$\frac{|w(t) * s(t) * \alpha h(t - 2R/c)|^2}{E[|w(t) * s(t) * \underline{c}(t) + w(t) * n(t)|^2]} \Big|_{t=t_M} \quad (3.2)$$

where an expectation is required in the denominator because the interference components are random. The output signal-to-interference-plus-noise ratio (SINR) in Eq. (3.2) can be written in the frequency domain as

$$SINR = \frac{\left| \alpha \int_{-\infty}^{\infty} W(F) S(F) H(F) e^{-j4\pi RF/c} e^{j2\pi F t_M} dF \right|^2}{\int_{-\infty}^{\infty} |W(F)|^2 [|S(F)|^2 \Psi_{cc}(F) + \Psi_{nn}(F)] dF}. \quad (3.3)$$

where  $\Psi_{cc}(F)$  was defined previously and  $\Psi_{nn}(F)$  is the PSD of the additive receiver noise. Because of the stationarity assumptions, the denominator of Eq. (3.3) does not depend on the sample time  $t_M$ . Defining  $\Psi(F) = |S(F)|^2 \Psi_{cc}(F) + \Psi_{nn}(F)$  and noting that Eq. (3.3) is essentially the same as the standard maximization problem used to derive the standard matched filter [33], the maximum achievable SINR is

$$SINR_{\max} = |\alpha|^2 \int_{-\infty}^{\infty} \frac{|S(F)H(F)|^2}{\Psi(F)} dF. \quad (3.4)$$

This maximum achievable SINR is derived using the Schwarz inequality and occurs only when the receive filter is

$$W(F) \propto \frac{S(F)H(F)e^{j2\pi F(t_M - 2R/c)}}{\Psi(F)}. \quad (3.5)$$

Eq. (3.4) gives the maximum SINR for a given waveform, and Eq. (3.5) describes the transfer function of the receive filter that will achieve this SINR, but we still need to derive the waveform that optimizes SINR. We note that the optimum receive filter does not depend on the received signal strength as controlled by the coefficient  $\alpha$ . Using the method of Lagrange multipliers, SINR can be optimized through the energy spectrum of the transmit waveform. The Lagrange multiplier technique yields [32]

$$|S(F)|^2 = \max \left[ 0, \frac{\sqrt{|H(F)|^2 \Psi_{nn}(F)}}{\Psi_{cc}(F)} \left( A - \sqrt{\frac{\Psi_{nn}(F)}{|H(F)|^2}} \right) \right]. \quad (3.6)$$

According to Eq. (3.6), the waveform spectrum that maximizes received SNR at a particular sample time (which can be chosen arbitrarily) due to a target with transfer function  $H(F)$  is obtained by evaluating various ratios of the target, signal-dependent interference, and additive interference strengths at every frequency within the allowable spectral band. For certain values of  $A$ , the term in parenthesis will be negative and the optimum waveform has zero energy at that frequency. The constant  $A$  must be selected in order to satisfy an energy constraint on the optimum waveform (otherwise, SINR can be increased arbitrarily by simply transmitting more power). Assuming realistic constraints on the maximum peak power,  $P_t$ , that the radar can produce and a maximum allowable waveform duration  $T$ , the product of these two constraints yields the maximum energy that can be allocated in a single waveform. Hence, the constant  $A$  must be set such that

$$\int_{-\infty}^{\infty} |S(F)|^2 dF = P_t T = E_s. \quad (3.7)$$

To find the optimum waveform spectrum, one must perform a one-dimensional search over the constant  $A$  until the energy constraint is met.

It is interesting to note in Eq. (3.6) that the PSD of the clutter,  $\Psi_{cc}(F)$ , is in the denominator of the term that sits in front of the term that depends on the constant  $A$ . The presence of the clutter PSD means that every frequency will eventually saturate as the interference at that frequency becomes dominated by clutter rather than additive noise. When this occurs, adding more energy to that frequency does not result in an increase in SNR, because the clutter power will increase proportionately to the signal power. Eventually, all frequencies will become saturated in this way, and it becomes impossible to increase SINR without expanding the waveform's instantaneous bandwidth.

It is also critically important to note that the previous design equations only define the waveform's energy spectrum but not its specific shape. In particular, the phase of the frequency response of the waveform is not defined; therefore it is a free parameter that can be used to achieve other design goals while still maximizing

SINR. For example, it may be possible to design the specific waveform modulation to possess low cross-correlation with a second waveform being transmitted in a MIMO strategy.

In the absence of an external clutter response, the theory for finding the optimum waveform reduces to the theory presented in Ref. [10]. Specifically, the SINR to be optimized over the waveform spectrum becomes

$$SINR_{\max} = |\alpha|^2 \int_{-\infty}^{\infty} \frac{|S(F)H(F)|^2}{\Psi_{nn}(F)} dF. \quad (3.8)$$

Rather than optimize via Lagrange multipliers, the optimization can be expressed as a solution to the integral equation

$$\lambda s(t) = \int_{-0.5T}^{0.5T} s(\gamma)K(t-\gamma)d\gamma \quad (3.9)$$

where the optimum waveform is the eigenfunction corresponding to the largest eigenvalue,  $\lambda$ , of the kernel [10,32]

$$K(t) = \int_{-\infty}^{\infty} \frac{|H(F)|^2}{\Psi_{nn}(F)} e^{j2\pi F t} dF. \quad (3.10)$$

### 3.4.1.2 Random Target Impulse Response

Next, we consider a problem formulation where the target impulse response is modeled as a random process. We derive optimum waveform spectra using both an expected SNR metric and a MI metric.

To begin, we first define a target impulse response that is a realization of a wide-sense stationary, zero-mean, complex Gaussian random process. This particular target model is clearly nonphysical, as any practical target will have a finite physical size and cannot produce an infinitely long impulse response with identical statistics over all time. Nevertheless, we can use this assumption to develop a foundation for SNR-optimizing waveforms for random targets, and then consider the practical implications of a finite-duration target. Let the infinite-duration, stationary random target be denoted as  $g(t)$  having a PSD  $\Psi_{gg}(F)$ . Using the model of Fig. 3.1 and swapping the impulse response  $h(t)$  for the random impulse response  $g(t)$ , target, the PSD of the output signal contribution is

$$|S(F)|^2 \Psi_{gg}(F)$$

and the PSD of the output interference component is

$$|S(F)|^2 \Psi_{cc}(F) + \Psi_{nn}(F).$$

In Ref. [34] and others, a *local SNR* has been defined for random processes, which is the ratio of the signal PSD to the interference PSD over a narrow band of frequencies. Taking the ratio of the earlier two PSDs, we define an SINR spectral density as

$$\Psi_{SINR}(F) = \frac{|W(F)|^2 |S(F)|^2 \Psi_{gg}(F)}{|W(F)|^2 (|S(F)|^2 \Psi_{cc}(F) + \Psi_{nn}(F))}. \quad (3.11)$$

Different frequency components of a stationary Gaussian random process are statistically independent. Therefore unlike for the deterministic-target case, there is nothing that the filter  $W(F)$  can contribute in terms of causing different frequency components to coherently combine over the integration. In other words, for a true random target, the receive filter cannot provide any processing gain, and we instead assume that the receive filter is an ideal filter over the operating band  $B$  of the radar system. Therefore within the radar's operating band, the filter response cancels, and outside of the operating band, no signal or interference are received. Just as we would integrate a PSD to compute total power, here we integrate the SINR spectral density over the radar band to achieve a total SINR. The total expected SINR for the random target becomes

$$SINR = \int_B \frac{|S(F)|^2 \Psi_{gg}(F)}{(|S(F)|^2 \Psi_{cc}(F) + \Psi_{nn}(F))} dF \quad (3.12)$$

where the integral is now over the radar's bandwidth.

We now desire to maximize Eq. (3.12) subject to an energy constraint on the waveform, as given in Eq. (3.7). This optimization problem is once again solved using Lagrange multipliers. In fact, it is the same optimization problem as in Eq. (3.4) except that the squared magnitude of the deterministic target's transfer function has been replaced by the random target's PSD. Therefore the optimum waveform spectrum is

$$|S(F)|^2 = \max \left[ 0, \frac{\sqrt{\Psi_{gg}(F)\Psi_{nn}(F)}}{\Psi_{cc}(F)} \left( A - \sqrt{\frac{\Psi_{nn}(F)}{\Psi_{gg}(F)}} \right) \right], \quad (3.13)$$

which is identical in form and behavior to Eq. (3.6) except for the different characterization of the target.

The major difficulty with the waveform specified in Eq. (3.13) is that it is built on the assumption of a stationary random target, which cannot be true for physical targets. In Ref. [32], a method was developed to modify this waveform design for finite-duration targets. The method makes an alternative assumption, namely, that the random target impulse response is statistically stationary within some finite interval, and zero outside of that interval. Therefore let a random target impulse response  $h(t)$  be formed by multiplying the WSS target  $g(t)$  with a rectangular window of duration  $T_h$ . Outside of the window's interval of length  $T_h$ , the target impulse response is zeroed out, but inside the interval, the target's impulse response retains its Gaussian and stationary properties. Because any realization of the

finite-duration random target will have finite energy, its Fourier transform will exist and the expected energy of a target realization will be

$$E_H = E \left[ \int_{-\infty}^{\infty} |H(F)|^2 dF \right] = \int_{-\infty}^{\infty} E [|H(F)|^2] dF. \quad (3.14)$$

Based on Eq. (3.14), we can interpret  $E [|H(F)|^2]$  as the target's energy spectral density, or ESD, which we intend to be a parallel concept to the PSD of a stationary random process. In case the target impulse response has a nonzero mean spectrum  $\mu_H(F)$ , we can also define an energy spectral variance (ESV) according to

$$\sigma_H^2(F) = E [|H(F) - \mu_H(F)|^2], \quad (3.15)$$

but for our current assumption of a zero-mean target we have  $\sigma_H^2(F) = E [|H(F)|^2]$ , such that the ESD and ESV are equal.

The waveform definition in Eq. (3.13) requires a description of the target's average power spectrum, not its average energy spectrum because it was derived under the assumption of an infinite-energy stationary random target. To resolve this disconnect, we choose to perform a time averaging of the target's ESV to obtain a power quantity. The target impulse response was defined to have a duration of  $T_h$ ; therefore we define the target's power spectral variance (PSV) as

$$\xi_H^2(F) = \frac{E [|H(F)|^2]}{T_h}. \quad (3.16)$$

The finite-duration target's PSV is obtained by taking the expected energy of every spectral component, and averaging over the target's time duration. Substituting Eq. (3.16) into Eq. (3.14), the target's expected energy is

$$E_H = T_h \int_{-\infty}^{\infty} \xi_H^2(F) dF, \quad (3.17)$$

and we see that  $\xi_H^2(F)$  can be appropriately interpreted as a power density.

Because the target impulse response is finite duration, so is the signal component of the filter output  $z(t)$ . The SINR expression of Eq. (3.12) has the term  $|S(F)|^2 \Psi_{gg}(F)$ , which is the PSD of an infinite-duration signal output. Using the earlier techniques, we can write an alternate expression for the signal output due to a finite-duration random target as  $|S(F)|^2 \sigma_H^2(F)$ , which we should interpret as the output ESD. To obtain a power quantity, we again perform time averaging, but the duration of the output signal is the result of convolving a target response of duration  $T_h$  with a waveform of duration  $T$ . Therefore the duration of the signal at the output of the filter is  $T_z = T + T_h$ , and the output PSV is

$$\xi_Z^2(F) = \frac{|S(F)|^2 \sigma_H^2(F)}{T_z} \cdot \frac{T_h}{T_h} = \frac{\epsilon |S(F)|^2 \sigma_H^2(F)}{T_h} = \epsilon |S(F)|^2 \xi_H^2(F) \quad (3.18)$$

where  $\epsilon = T_h/T_z = T_h/(T + T_h)$  contains the ratio of the target duration to the overall output duration. This ratio accounts for the fact that the finite-duration target causes

the output PSV of the waveform-target interaction to vary with time. Substituting Eq. (3.18) for the numerator of the integrand in Eq. (3.12) and performing the optimization, we have

$$|S(F)|^2 = \max \left[ 0, \frac{\sqrt{\xi_H^2(F)\Psi_{nn}(F)}}{\Psi_{cc}(F)} \left( A - \sqrt{\frac{\Psi_{nn}(F)}{\xi_H^2(F)}} \right) \right]. \quad (3.19)$$

If we take the limit as the target duration becomes infinitely long, then  $\xi_H^2(F)$  approaches  $\Psi_{gg}(F)$  and Eq. (3.19) converges back to Eq. (3.13).

The expression in Eq. (3.19) provides an approximation to the optimized spectrum from the standpoint of maximizing expected SNR from a Gaussian ensemble of finite-duration random targets. In other applications, MI is often deemed to be a better optimization metric. Once again, we will need to take a theoretical result for stationary Gaussian random processes and then make an adjustment for finite-duration targets.

Let us try to maximize the MI between the output signal  $z(t)$  and a stationary complex-Gaussian, zero-mean target impulse response  $g(t)$ . The MI will be maximized over the transmit waveform spectrum,  $S(F)$ . Based on stationary Gaussian random process, the rate of MI per unit time is

$$\frac{d}{dt}[I(z(t);g(t)|s(t))] = \int_B \ln \left[ 1 + \frac{|S(F)|^2 \Psi_{gg}(F)}{|S(F)|^2 \Psi_{cc}(F) + \Psi_{nn}(F)} \right] dF. \quad (3.20)$$

The expression in Eq. (3.20) is consistent with the well-known expression for MI of a Gaussian channel [35]. When a complex Gaussian random variable is the input to a channel, and the output of the channel is the input signal plus complex additive white Gaussian noise (AWGN), then the MI between output and input is  $\ln(1+SNR)$  where the  $SNR$  is defined as the ratio of the variance of the input random variable to the variance of the additive noise. Eq. (3.20) mimics this expression, but by acknowledging again that every infinitesimally narrow frequency component of a Gaussian random process is statistically independent, the total information rate is obtained by integrating over all frequencies within the bandwidth of the random process.

Because a stationary Gaussian random process exists for all time, the entropy of such a random process is infinite, and the MI for an infinitely long observation interval is also infinite. As such, Eq. (3.20) is not directly useful to us, and we take the same approach as previously by substituting the time-averaged output PSV for the output PSD, yielding

$$\frac{d}{dt}[I(z(t);h(t)|s(t))] \approx \int_B \ln \left[ 1 + \frac{\epsilon |S(F)|^2 \xi_H^2(F)}{|S(F)|^2 \Psi_{cc}(F) + \Psi_{nn}(F)} \right] dF. \quad (3.21)$$

If Eq. (3.21) is an information rate, then the total MI obtained during an observation interval of duration  $T_z$  is

$$I(z(t); h(t)|s(t)) \approx T_z \int_B \ln \left[ 1 + \frac{\epsilon |S(F)|^2 \xi_H^2(F)}{|S(F)|^2 \Psi_{cc}(F) + \Psi_{nn}(F)} \right] dF. \quad (3.22)$$

In reality, the information rate is not constant due to the beginning and ending transition periods of the convolution between target impulse response and waveform. However, the effect of the time duration ratio  $\epsilon$  is to at least approximately account for this effect.

Now that we have taken the theoretically correct expression for stationary Gaussian random processes and made modifications to account for finite-duration targets, we now need to maximize the information in Eq. (3.22) over the waveform spectrum  $|S(F)|^2$ , subject to the waveform energy constraint. Owing to the concavity of the integrand of Eq. (3.22), the Lagrange multiplier technique yields an optimum waveform spectrum defined by

$$|S(F)|^2 = \max \left[ 0, -B(F) + \sqrt{B^2(F) + C(F)(A - D(F))} \right] \quad (3.23)$$

where

$$B(F) = \frac{\Psi_{nn}(F)(2\Psi_{cc}(F) + \epsilon\xi_H^2(F))}{2\Psi_{cc}(F)(\Psi_{cc}(F) + \epsilon\xi_H^2(F))}, \quad (3.24)$$

$$C(F) = \frac{\epsilon\Psi_{nn}(F)\xi_H^2(F)}{\Psi_{cc}(F)(\Psi_{cc}(F) + \epsilon\xi_H^2(F))}, \quad (3.25)$$

and

$$D(F) = \frac{\Psi_{nn}(F)}{\epsilon\xi_H^2(F)}. \quad (3.26)$$

As in the previous optimizations, the waveform spectrum is found via a one-dimensional search for the constant  $A$  that results in the energy constraint being satisfied. In the case where the external signal-dependent interference (e.g., ground clutter) is nonexistent, then  $\Psi_{cc}(F) = 0$ , and the waveform solution becomes

$$|S(F)|^2 = \max [0, A - D(F)]. \quad (3.27)$$

The previous equations describe custom-designed waveform spectra that can be used to enhance SNR and/or information extraction. They are based on Gaussian assumptions, which will not be entirely accurate for many applications. Yet application of the previous design equations to target detection and target recognition applications has yielded improved performance over that achieved by standard waveforms with approximately flat spectrum over a band of interest. Similar equations have even been derived in two spatial frequency dimensions for optimized transmit beamforming [36]. The general strategy is to define a quantity of interest, such as target presence/absence or target type, to compute a spectral variance in the appropriate dimensions over the hypotheses of interest, and then to use the design equations to optimize the radar's transmit illumination over certain

parameters or design variables. Application examples will be demonstrated later in this chapter.

#### 3.4.1.3 Waveform Shape and Constant Modulus Constraints

As mentioned earlier, the previous equations define the optimal waveform's energy spectrum but do not specify the waveform's phase function. Therefore the specific waveform shape is not completely specified, and some design flexibility remains. There are various requirements or constraints that one may wish to place on the waveform, but one common constraint is for the waveform's time-domain envelope to have a constant amplitude. This so-called constant modulus constraint [37–39] is very important, as physical radar transmitters have a peak power capability that cannot be exceeded. Therefore the peak amplitude of the time-domain waveform must be scaled to prevent saturation of the transmitter, and any resulting time spent at less than peak amplitude implies that the radar is transmitting at less than peak power. Considering that the energy constraint in Eq. (3.7) assumes a constant power over the duration of the waveform, a time-varying envelope combined with a peak power constraint means that the maximum energy cannot be achieved, and the optimized spectrum loses its usefulness.

Several researchers have studied the problem of designing a constant modulus waveform with a specific energy spectrum [37–39], with alternating projections in the time and frequency domain being one of the more common approaches. The waveform that results from these techniques will be constant modulus but will not match the intended waveform spectrum exactly. For example, any practical, finite-duration waveform cannot produce finite contiguous frequency bands with zero energy and must have frequency domain sidelobes. Hence, we conclude that in many cases it will not be possible to produce the ideal waveform spectrum exactly, but that the spectrum can usually be approximated very closely. For a given application, it is important to understand the sensitivity of any desired performance gains to small errors in the ideal waveform spectrum, such that realistic waveform errors may or may not cancel the benefit intended through optimization. More detail on implementing constant modulus constraints is available in Ref. [31].

#### 3.4.2 SEQUENTIAL HYPOTHESIS TESTING

Much of the potential benefit of cognitive radar is based on optimal allocation of sensing resources. Properly designed waveforms can enhance SNR or improve separation between target classes, optimized beamsteering or beamshaping can be used to focus energy on those spatial locations where target presence or absence is most in doubt, and target tracks can be updated when necessary rather than on a fixed timeline. To serve this purpose, it is often useful to know when the radar has collected sufficient data for a reliable decision among competing hypotheses. Sequential hypothesis testing [14] provides just such a framework.

In the standard textbook formulation of statistical decision-making, the size of the data record is assumed to be fixed. Probability density functions are specified for the

data record under each of the potential hypothesis, and the decision-making procedure is derived from there. In sequential hypothesis testing, an additional decision option is provided. This additional option allows for the system to determine that insufficient data exist for making a decision of the required accuracy. Additional data are required, at which time the system updates the relevant likelihood functions and decides again whether to choose a hypothesis or to require more data. The procedure continues until the accumulated data support a final decision that meets the required accuracy. In many applications, this sequential decision-making is a perfect fit for the closed-loop nature of cognitive radar.

A variety of procedures exist for sequential testing of different simple and composite hypotheses, and for binary decision making versus multihypothesis testing. For binary testing, at each iteration of the decision-making procedure, a decision is made to either accept the hypothesis, reject the hypothesis, or to collect additional data. As soon as the hypothesis is either accepted or rejected, additional data are not needed and the data collection (for this particular hypothesis) ends. The termination criteria require specification of allowable error probabilities, which translate to thresholds on the likelihood ratio that must be met. The amount of data necessary to reach a decision threshold will vary from test to test; therefore the amount of data required is random and the best metric for performance is the expected amount of data required. In the terminology of sequential testing, the expected amount of data required is called the average sample number (ASN).

#### **3.4.2.1 Binary Sequential Hypothesis Testing**

To consider how a sequential hypothesis test works, consider the case of deciding between two simple hypotheses. Let the data collected during the  $k$ th iteration of the testing procedure be denoted as  $\mathbf{z}_k$ , and the pdf of the observed data for hypotheses #1 and #2 on the  $k$ th data collection be  $p_{1k}(\mathbf{z}_k)$  and  $p_{2k}(\mathbf{z}_k)$ . Let the prior probabilities of the two hypotheses be  $P_1$  and  $P_2$ . Although prior probabilities such as these are not normally defined in radar applications, prior information of some kind is usually available. For example, expected target densities or the likelihood of observing different target types could be known and turned into prior probability estimates. After the first data collection, the likelihood ratio is

$$\Lambda^1(\mathbf{z}_1) = \frac{p_{11}(\mathbf{z}_1)P_1}{p_{21}(\mathbf{z}_1)P_2}.$$

This likelihood ratio is compared to a pair of thresholds  $\eta_1$  and  $\eta_2$  where  $\eta_1 > \eta_2$ . If  $\Lambda^1 > \eta_1$ , then a decision is made for hypothesis #1, and the experiment is terminated. If  $\Lambda^1 < \eta_2$ , then a decision is made for hypothesis #2, and the experiment is terminated. Otherwise, another round of data collection is performed. After the second data collection, the resulting SPRT is

$$\Lambda^2(\mathbf{z}_1, \mathbf{z}_2) = \frac{p_1(\mathbf{z}_1, \mathbf{z}_2)P_1}{p_2(\mathbf{z}_1, \mathbf{z}_2)P_2} = \frac{p_{11}(\mathbf{z}_1)p_{12}(\mathbf{z}_2)P_1}{p_{21}(\mathbf{z}_1)p_{22}(\mathbf{z}_2)P_2} \quad (3.28)$$

where the second equality assumes that the two data collections are statistically independent (they do not have to be independent, but the sequential likelihood ratio requires computing successive joint probability densities). Also, the expression in Eq. (3.28) allows for the pdfs to vary from data collection to data collection. A more narrowly defined case could define the pdfs to be identical over all data collections, but the closed-loop and adaptive-transmit nature of cognitive radar implies that the form of the data, and hence their pdf, will likely vary during the experiment. The new likelihood ratio is again compared to the two thresholds, with the possible decisions being to select hypothesis #1, select hypothesis #2, or to continue the experiment. Eventually, assuming statistically independent observations, the likelihood ratio after  $k$  data collection is

$$\Lambda^k(\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_k) = \frac{p_{11}(\mathbf{z}_1)p_{12}(\mathbf{z}_2)\dots p_{1k}(\mathbf{z}_k)P_1}{p_{21}(\mathbf{z}_1)p_{22}(\mathbf{z}_2)\dots p_{2k}(\mathbf{z}_k)P_2}. \quad (3.29)$$

We say that the required number of data collections was  $K$ . We also note that for the statistically independent case, there is an equivalent sequential test using log-likelihoods, according to

$$\ln [\Lambda^k(\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_k)] = \ln \left[ \frac{p_{11}(\mathbf{z}_1)}{p_{21}(\mathbf{z}_1)} \right] + \ln \left[ \frac{p_{12}(\mathbf{z}_2)}{p_{22}(\mathbf{z}_2)} \right] + \dots + \ln \left[ \frac{p_{1k}(\mathbf{z}_k)}{p_{2k}(\mathbf{z}_k)} \right] + \ln \left[ \frac{P_1}{P_2} \right]. \quad (3.30)$$

In many cases, the log-likelihood sequential test will be simpler to compute and may be more numerically stable. In the log-likelihood form, the thresholds are  $\ln \eta_1$  and  $\ln \eta_2$ .

Next, we need to determine appropriate thresholds. If the two likelihoods comprising the likelihood ratio (i.e., the numerator and denominator of Eq. 3.29) are equal, then the likelihood ratio value is  $\Lambda^k = 1$ . We can easily see that we must choose  $\eta_1 > 1$  and  $\eta_2 < 1$ . Moreover, the larger we set  $\eta_1$ , the less likely we are to incorrectly choose hypothesis #1 when hypothesis #2 is true, and the smaller we set  $\eta_2$ , the less likely we are to incorrectly choose hypothesis #2 when hypothesis #1 is true. The procedure for setting the thresholds based on maximum allowable error rates is given in Ref. [14]. Let the error rate for choosing hypothesis #1 when hypothesis #2 is true be  $\beta_1$ , and let  $\beta_2$  be the error rate for the opposite situation. According to Ref. [14], for a test between two simple hypotheses, the inequalities

$$\eta_1 \leq \frac{1 - \beta_2}{\beta_1} \quad (3.31)$$

and

$$\eta_2 \geq \frac{\beta_2}{1 - \beta_1} \quad (3.32)$$

hold true. Rearranging, it can also be stated that

$$\beta_1 \leq \frac{1}{\eta_1} \quad (3.33)$$

and

$$\beta_2 \leq \eta_2. \quad (3.34)$$

In other words, the strength of the test, as defined by maximum error rates on both types of errors, can be controlled by selecting  $\eta_1$  sufficiently large and  $\eta_2$  sufficiently small.

Now that we have specified the procedure for the sequential hypothesis test and set appropriate thresholds to ensure that we meet our bounds on error rates, we must consider the value in implementing such a test. Why implement a sequential test rather than a standard likelihood ratio test based on a fixed number of observations? First, the sequential testing procedure naturally aligns with the iterative, closed-loop nature of cognitive radar. As data are received, the cognitive radar system can quantify its learning in a Bayesian sense by updating the probabilities on various scenarios of interest and by updating pdfs on parameters of interest. These updated probabilities and pdfs can then be used to optimize the form of the next illumination, which then results in new likelihoods to fold into the next iteration of the sequential test and so on. Second, we find that the sequential hypothesis test requires fewer measurements on average than a fixed test having the same error rates. So the sequential testing procedure is not only consistent with closed-loop adaptive sensing, but it also saves resources by terminating the decision-making procedure as soon as sufficient confidence is obtained. Detailed analysis of the ASN under various conditions is given in Ref. [14] and in the subsequent literature. Fundamental examples of the benefits of combining adaptive sensing with sequential hypothesis testing are given later in the chapter.

In considering the  $k$ -measurement SPRT of Eq. (3.29), we see that at each iteration, the sequential test compares a ratio of posterior probabilities to a set of thresholds. The prior probabilities quantify the likelihood of each hypothesis prior to any data collections. After the first measurement, these probabilities are updated via Bayes' Theorem to obtain posterior probabilities. For example, after the first data collection, the posterior probability of hypothesis #1 is

$$\Pr\{\mathcal{H}_1 | \mathbf{z}_1\} = \frac{p(\mathbf{z}_1 | \mathcal{H}_1)}{p(\mathbf{z}_1)} P_1 = \frac{p_{11}(\mathbf{z}_1)}{p(\mathbf{z}_1)} P_1. \quad (3.35)$$

The normalizing term  $p(\mathbf{z}_1)$  is difficult to compute, but is not necessary because the posterior probability of the second hypothesis has the same factor; hence, they cancel out. After each data collection, the sequential test can be considered as an update to the probabilities of each hypothesis being true. If the probability of one hypothesis becomes sufficiently strong compared to the other, then the test is terminated, and the definition of *sufficiently strong* is quantified by the relationship of the testing thresholds to the desired error rates.

### 3.4.2.2 Sequential Testing with Multiple Hypotheses

The sequential testing thresholds in Eqs. (3.31) and (3.32) depend on the error rates for both error types. Moreover, only two hypotheses at a time can be compared in a single ratio. These factors make a direct extension of sequential hypothesis testing to

more than two hypotheses a challenge in terms of determining the proper likelihoods to use and the thresholds to which they should be compared. One logical extension is to simply track a parallel set of sequential tests covering all possible pairs of competing hypotheses and to continue the experiment until one hypothesis is favored with acceptable confidence over all other hypotheses. This straightforward extension to the binary sequential hypothesis test is, unfortunately, suboptimum in the sense of minimizing the expected sample number for a set of desired error rates [18]. But it is simple to implement and has proven to be useful [6] despite its suboptimal nature.

This particular form of multihypothesis sequential testing is called the multihypothesis SPRT [16]. Let  $\beta_{i,j}$  for  $i \neq j$  be the allowable error rate for selecting hypothesis  $\mathcal{H}_j$  when hypothesis  $\mathcal{H}_i$  is true. After the  $k$ th data collection, we form a full set of pairwise SPRTs according to

$$\Lambda_{i,j}^k(\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_k) = \frac{p_{i1}(\mathbf{z}_1)p_{i2}(\mathbf{z}_2)}{p_{j1}(\mathbf{z}_1)p_{j2}(\mathbf{z}_2)} \cdots \frac{p_{ik}(\mathbf{z}_k)P_i}{p_{jk}(\mathbf{z}_k)P_j}. \quad (3.36)$$

This  $(i,j)$ th element of the multihypothesis SPRT is compared to determine if it exceeds the threshold

$$\eta_{i,j} = \frac{1 - \beta_{i,j}}{\beta_{j,i}}. \quad (3.37)$$

Instead of also comparing Eq. (3.36) to see if it falls below a low threshold, we instead also compute  $\Lambda_{j,i}^k$  and compare it to a threshold  $\eta_{j,i}$ . In this way, we can define a stopping criterion that selects hypothesis  $\mathcal{H}_i$  when the condition

$$\Lambda_{i,j}^k > \eta_{i,j}; \quad \forall j \neq i \quad (3.38)$$

is met for some hypothesis  $\mathcal{H}_i$ .

### 3.4.3 PARTIALLY OBSERVABLE MARKOV DECISION PROCESS

Another technique that has been effectively used in initial studies of cognitive radar is the use of partially observable Markov decision processes (POMDP) [40]. When applied to cognitive radar, POMDPs are another technique for maintaining a probabilistic representation of various beliefs about the state of the RF environment. The beliefs can be updated with measurements, and more important, the POMDP can be used to maximize the accumulation of expected future rewards based on actions taken. Actions are performed sequentially, that is, at each time step, a decision is made regarding the next action to take based on an optimization of expected reward over some future time horizon. Therefore POMDPs can be used as an optimization mechanism to control sensing actions and parameters, although most POMDPs are too complex to solve exactly and require approximate solutions. The possible underlying states of the RF environment are modeled as a Markov decision process (MDP), which means that the underlying state of the environment can be dynamically changing over time according to defined state transitions.

We will see shortly that, even for simple scenarios as in the canonical examples of the next section, the ability to probabilistically represent possible states of the world, and to convert those probabilistic states into action, is an underlying core component of cognitive radar. Just as in sequential hypothesis testing as described previously, POMDPs allow for tracking of probabilities on different hypotheses over time. However, sequential hypothesis testing requires the hypotheses to be static and the measurement distribution to be consistent over an experiment. In contrast, POMDPs enable dynamic and potentially more complex scenarios through the more general use of state spaces and transition probabilities. Due to its MDP-based formulation, the probability density function of the state space at any given time is conditioned on and completely captures information gained through prior data.

In Ref. [41], a POMDP framework was used as the framework for a cognitive tracking algorithm that demonstrated the ability of *anticipation*, which was defined as the ability of the cognitive radar system to select optimal actions based on future expectations of the state space rather than on the current expectation of the state space. In particular, track update rate was optimized using two different methods: (1) based on optimization using only the current state, which maximizes short-term reward, and (2) based on POMDP-enabled calculation of future reward over a future sequence of actions. When presented with a scenario where the target would be occluded for several seconds, the POMDP-based optimization was shown to significantly increase its update rate prior to the occlusion, in order to sharpen the track prior to the window of time where the target would be unobservable. This anticipatory sharpening allowed the tracker to pick up the target again after the occlusion. In contrast, the optimization procedure based on only the current belief state failed to anticipate the occlusion, did not sharpen the track prior to the occlusion, and lost track during the occlusion. POMDPs have also been used for adaptive waveform selection in Refs. [42,43].

It is self-apparent that future cognitive radar systems should be able to anticipate changes in the sensing environment that may impact performance. POMDPs have been shown to have potential in this regard and may become critical components of the cognitive radar frameworks that develop in the upcoming years.

---

### 3.5 CANONICAL EXAMPLES

The earlier sections have presented some techniques that can be useful for cognitive radar, along with a few comments about the desire to maximize radar performance through adaptive transmission and feedback from the receiver to the transmitter. In this section, we wish to make the procedures and potential benefits more concrete by applying several of the techniques to several fundamental, textbook-type problems. We consider problems such as simple detection of a known signal, detection with a nuisance parameter, and multiple detection decisions in parallel. We also consider the issue of dealing with different types

of radar exploitation tasks, such as detection and parameter estimation, which have different types of information metrics.

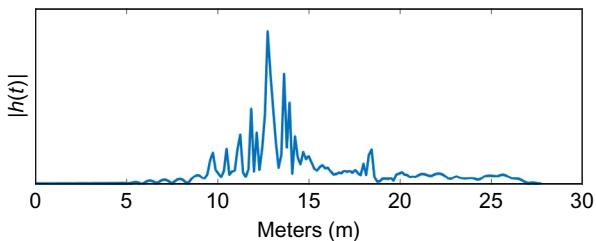
### 3.5.1 DETECTION OF A TARGET WITH KNOWN IMPULSE RESPONSE

To begin, we consider detection of a target with a known impulse response, or equivalently, a known target transfer function. The target is illuminated by a waveform, resulting in a known received waveform that is corrupted by AWGN. For a detection problem involving a deterministic quantity embedded in Gaussian noise, detection performance is perfectly correlated to SNR, so we will use the earlier technique for maximizing SNR through matched waveform design. We will look at detection performance of the optimized-SNR waveform compared to a standard, flat-spectrum waveform. However, we will also look at information gained with respect to the detection decision, and whether the diminishing information gains at high SNR suggest that some of the radar's resources would have been better applied toward other sensing tasks.

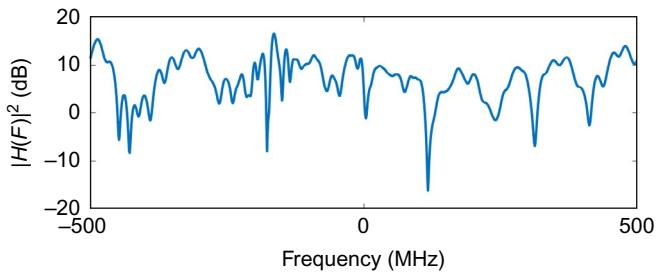
#### 3.5.1.1 Waveform Design

A real target of interest is a complicated structure comprising many different scattering points and surfaces. Reflections from these scatterers, including multiple-bounce reflections, all have different propagation path lengths and scattering strength. The coherent superposition of these scattering components produces complicated scattering behavior that varies with frequency as the phases combine more or less constructively or destructively. Scattering points or surfaces that have path length differences too short to be resolved appear as a single scatterer with RCS that varies over frequency and radar aspect angle. When the radar achieves sufficient resolution that the different path lengths can be resolved, then we say that the target becomes *extended*, meaning the target response does not all fall within a single resolution cell. Of course, assuming perfect knowledge of the target impulse response is unrealistic in most applications; for example, small shifts on the scale of a fraction of a wavelength change the overall phase of the target response. But the deterministic case can provide some useful initial insights despite its over-optimistic assumptions.

Figs. 3.2 and 3.3 show samples of a target impulse response and target frequency response, respectively. The responses were obtained through electromagnetic modeling of a CAD model of a small fighter jet using the finite-difference time-domain technique. The modeling geometry was approximately from the side of the aircraft, such that the range dimension observed by the radar was just under 10 m from wingtip to wingtip. The resulting frequency response, as seen in Fig. 3.3, reveals the frequency selective behavior whereby the scattering combines more constructively at some frequencies than at others. The target responses were simulated over a 1-GHz frequency band, and then translated to a complex baseband representation.

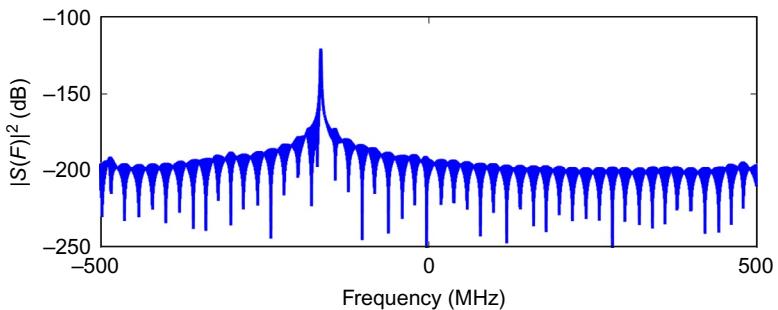
**FIG. 3.2**

Magnitude of the target impulse response for a fighter jet as observed from the side of the jet. Modeled via EM software (XFdtd).

**FIG. 3.3**

Transfer function of a small fighter jet corresponding to the impulse response seen in Fig. 3.2.

The ability of the radar to resolve different parts of the target depends on having sufficient bandwidth to observe this fluctuating frequency-domain behavior. The bandwidth at which the fluctuating behavior becomes apparent is also the bandwidth at which the range resolution becomes smaller than the physical target size. However, resolution is important for forming an image or range profile, not for detection, and Fig. 3.4 shows that the spectrum of the SNR-maximizing waveform is actually very narrowband. The optimized waveform shown in Fig. 3.4 was restricted to a time duration of 1  $\mu$ s, peak transmit power of 1 W, and to the same 1-GHz operating band that is displayed in Fig. 3.3. Under these limitations, the optimized waveform is a narrow tone centered at approximately -165 MHz, which corresponds to the peak location of the target frequency response. This behavior is intuitive and consistent with earlier published work [11,12] on SNR-optimizing waveforms—the strongest received signal will be obtained when the transmit energy is focused into the narrow band with the highest target strength (e.g., largest RCS). In this particular case, which does not have signal-dependent interference, the shape of the ideal waveform spectrum does not vary with transmit power level.

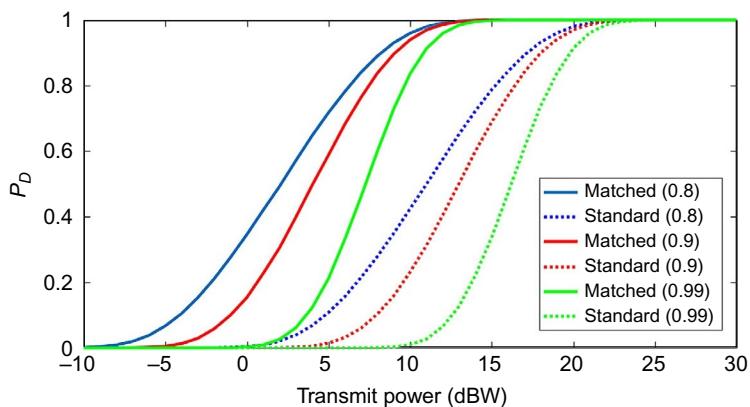
**FIG. 3.4**

Spectrum of the SNR-optimized waveform matched to target spectrum in Fig. 3.3.

### 3.5.1.2 Detection Performance

Next, we consider detection performance and information obtained with the ideal illumination. The waveform-target interaction is modeled as a linear system via convolution, and the simulated received signal is scaled by  $1/R^2$  where  $R = 10$  km is the assumed target range. Complex AWGN with flat PSD of  $N_0 = 1\text{e}{-20}$  is added to the received signal, and the noisy result is passed through a matched filter and sampled at the time of the peak signal response. Because the waveform, target impulse response, matched filter, and receiver noise power are known, the signal component at the sample time is known, and the distribution of this sample is Gaussian with known mean and variance.

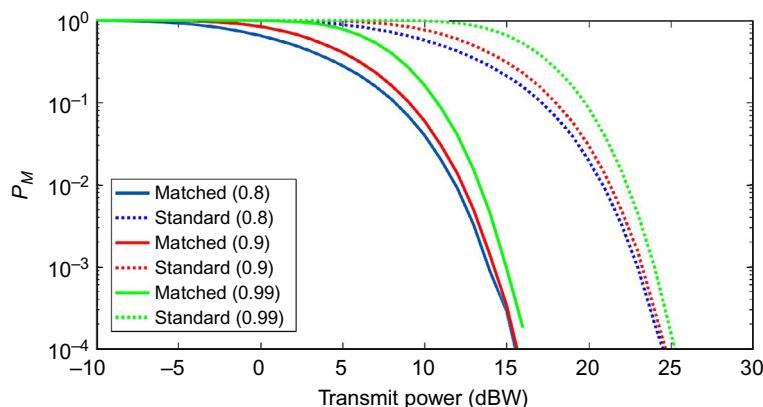
Fig. 3.5 shows target detection performance for different threshold levels and varying transmit power (the optimized waveform was calculated at 1 W, or

**FIG. 3.5**

Detection performance comparison for a matched waveform versus an unmodulated, standard 15-MHz waveform.

0 dBW, and scaled as necessary to obtain each data point). In traditional approaches, it is assumed that the prior probability of target presence cannot be defined or accurately determined. Therefore thresholds are usually set to achieve a fixed probability of false alarm according to the Neyman-Pearson Theorem [44]. In cognitive radar, however, in order to achieve closed-loop adaptivity and improved performance, it is important to be able to quantify knowledge and learning in some way that can be acted upon in future measurements. In a practical application, the system's history, outside circumstances, and other side information should inform the prior probability, and clearly an important challenge in cognitive radar is to characterize and quantify prior knowledge. For our purposes here, we consider a case where the prior probability of target presence before collecting any data is assigned to be  $P_1 = 0.5$ . Upon collecting a measurement, the probability of target presence can be updated via Bayes' Theorem (as in Eq. 3.35) to obtain a posterior probability. In a sequential testing procedure, the updated probability could be compared to upper and lower thresholds to determine whether to make a final decision for either hypothesis or to continue collecting data.

In this example, we simply compare the posterior probability to a threshold after a single measurement and immediately make a decision. Three different posterior probability thresholds are considered: 0.8, 0.9, and 0.99. The lowest threshold of 0.8 yields the highest detection probably but would also yield higher false detections. For this particular case, the matched waveform yields approximately 9 dB improvement in SNR, which corresponds to the approximate shift between curves for the two different waveforms at the same threshold value. This shift is also seen in Fig. 3.6, which shows the probability of miss for the same simulation. Probability of miss, as shown in semilog scale in Fig. 3.6, reveals interesting behavior at high SNR. We see that at high SNR, all curves approach a similar asymptotically linear behavior with



**FIG. 3.6**

Probability of miss comparison for a matched waveform versus an unmodulated, standard 15-MHz waveform.

the same slope. This slope quantifies the fixed rate of improvement in the asymptotic region of miss probability for every dB increase in transmit power. The 9-dB gain from waveform optimization shows as a shift in the asymptotic regions of the two sets of curves.

### 3.5.1.3 Information Gained

From Fig. 3.6, we see that as the radar system continues to apply more energy toward detection of this target, the probability of miss continues to decrease. However, from the perspective of information gained, once the asymptotic region is reached, the returns per dB of transmit power are diminished. To demonstrate this fact, we can compute the MI provided by the measurement. Prior to measurement, the prior probability of target presence is 0.5, yielding the maximum possible entropy of 1 bit for a binary decision or random variable. After collecting a measurement, the posterior probability of target presence can be used to compute a change of entropy, which is the information gained. Averaging this calculation over thousands of trials yields the MI performance versus transmit power shown in Fig. 3.7. The MI in Fig. 3.7 has been called the task-specific information (TSI) [45,46], as it is the MI specifically related to the exploitation task of target detection. TSI is not directly related to the entropy of the actual data, which depends on noise power and for some detection problems may depend on nuisance parameters. Instead, we see that for the optimized waveform, the detection-specific information gained flattens out at transmit power levels above 10 dB. So although Fig. 3.6 shows continued reductions in probability of miss, from a cognitive radar perspective we ask the question of whether power levels above approximately 12 dB are being efficiently used, or whether those sensing resources should be directed toward a different task. Physically speaking, peak transmit power probably will not be reduced, but energy on target could be

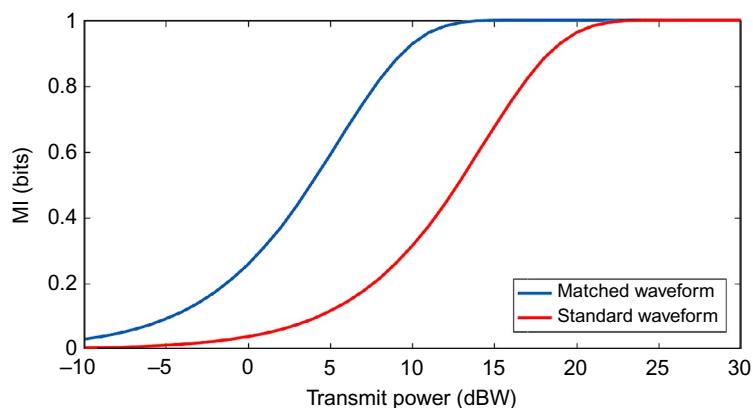


FIG. 3.7

Detection-specific mutual information gained by the radar system versus transmit illumination power.

reduced by shortening the number of pulses in a CPI or by shifting energy to another narrow frequency band that resonates with a different target. The purpose here is to consider whether cognitive radar systems with feedback and adaptive transmitters might enable better use of resources in various applications.

### 3.5.2 DETECTING A KNOWN SIGNAL WITH A NUISANCE PARAMETER

Nuisance parameters, such as unknown target range, Doppler shift, or aspect angle have the potential to complicate the detection problem (or other radar task such as classification), and in general, detection performance in the presence of nuisance parameters is worse than in the known signal case. SNR is reduced in the presence of nuisance parameters due to the inability to exactly specify the matched filter.

In signal processing for detection of a single target, nuisance parameters are typically handled as a deterministic unknown or as a random signal parameter. When the unknown parameter's probability density function is either unknown or cannot be defined, then the generalized likelihood ratio test [44] can be used. In this method, the maximum likelihood estimate of the parameter is computed, and then this estimate is substituted into the likelihood ratio for detection. In the case of a random unknown parameter, the likelihood ratio can be defined in terms of the signal pdf conditioned on the random parameter, and then the parameter can be integrated out of the likelihood ratio using Bayes' rule.

In another canonical example of cognitive radar, we can treat one or more unknown parameters as random quantities with defined pdf(s), update the pdf(s) as data are obtained, and then update the optimizations of subsequent measurements in order to improve decision performance or convergence time. In the prior example, the transmit waveform was matched to a known target response in order to maximize SNR and detection performance. In that example, the benefit of cognitive radar was obtained through explicit matching of a waveform to a known target, but there was no additional closed-loop adaptation of the radar behavior. In the case of a nuisance parameter, we can demonstrate some additional features of cognitive radar. First, we show the benefits of carrying information over from one set of measurements to the next. Many radar systems perform processing on a data set and then make hard decisions on target presence or absence. If a detection statistic falls below the detection threshold, even by a small amount, then the target is declared "not present" and the detection statistic is thrown away. The next time data are collected on the same potential target, the detection process is repeated without regard to the results from previous data collections.

It would be better if detection statistics could be carried over from one data collection to the next, either as direct values or as updated probabilities of target presence. In this way, targets with moderate SNR that cannot be reliably detected on a single processing interval can be revealed over time by integrating over multiple processing intervals. Of course, the radar scene is dynamic, such that the geometry can change, targets can maneuver, and old information can become stale. Therefore any method that attempts to carry information forward over time must include

appropriate methods for dealing with the dynamic nature of the radar environment. One such class of algorithms that attempts to carry information over multiple observations in order to enable detection of weak targets is known as Track-Before-Detect (TkBD) [47–49]. Generally speaking, TkBD algorithms identify potential targets that cannot yet be declared as detections, but should be monitored over multiple observations to check for consistency that would indicate target presence. Computationally, this procedure can only be performed for a finite number of potential targets, so most TkBD algorithms perform a soft detection to identify key locations in the target parameter space that are most promising for target presence. Target dynamics are accommodated via motion models applied to the potential target state, by allowing the uncertainty of the potential target state to vary, and by resampling (in the case of TkBD based on particle filtering) the numerical density functions that represent potential targets. Some algorithms explicitly calculate and carry forward a probability of target presence.

### 3.5.2.1 Waveform Design Applied to Adaptive Beamshaping

To demonstrate the philosophy, consider an active phased-array radar system capable of transmitting various beam shapes in search of radar targets. For now, we assume that the radar system can create arbitrary transmit illumination patterns in a single plane (e.g., the azimuth direction). Of course, any real system will have practical limitations to its realizable transmit patterns that must be considered in any optimization of the transmit pattern, but these limitations are not considered here. On receive, we assume that the system operates as a multichannel receiver that digitizes the signal received on each element; therefore detection of a target in a particular direction is performed by digital beamforming on receive to maximize SNR in that direction. Because the receive beamforming is digital, we can treat the target's angle of arrival (AOA) as a nuisance parameter in the detection problem and steer the receive beam to an arbitrary number of AOAs.

Let  $F_s = (d/\lambda) \sin\theta$ ,  $F_s \in [-0.5, 0.5]$ , be the normalized spatial frequency associated with the potential target's AOA,  $\theta$ , when the array element spacing is  $d$  and the system's operating wavelength is  $\lambda$ . If the target is present, its AOA, or equivalently its spatial frequency, is unknown. Therefore there is a compromise to be made on any given transmission. The transmit beam can be focused and steered in a particular direction to maximize gain and SNR in that direction, but if the target is actually located in a different direction, the target will not be illuminated. On the other hand, the transmit beam can be spoiled to illuminate all directions and guarantee the target is illuminated, if present, but at the expense of the maximum gain in the transmit beam. Here, we demonstrate a potential cognitive, adaptive-transmit approach to mitigating this compromise.

In this simulated example, we assume a 16-element, uniformly distributed array. When fully focused, the array has a maximum gain of 30 dB. The maximum gain per element on receive is 18 dB (in practice, each element is more likely to be a subarray), the gain achieved through pulse compression is 23 dB, the operating frequency is 10 GHz, and the effective noise temperature of the individual receivers

is  $T_E = 725$  K. At this noise temperature and a signal bandwidth of 10 MHz, the receiver noise power per element is  $P_n = k_b T_E B = -130$  dBW where  $k_b$  is Boltzmann's constant. The transmit power is varied from  $-10$  to  $30$  dBW. The target, if present, is located at a range of 5 km and has an RCS of  $10 \text{ m}^2$ . Prior to the first transmission, the target location is assumed to be unknown and uniformly distributed between  $F_s = -0.4$  and  $F_s = 0.4$ . After each illumination and reception, the pdf of the AOA nuisance parameter is updated and, in the adaptive illumination case, used to optimize subsequent transmission patterns.

Let the prior pdf of the AOA nuisance parameter be  $p(F_s)$  and the complex,  $N$ -element array manifold vector that varies with AOA be  $\mathbf{a}(F_s)$ . The received measurements are modeled as

$$\mathbf{y} = \chi \alpha S(F_s) \mathbf{a}(F_s) + \mathbf{n}$$

where  $\alpha$  is an amplitude scaling factor determined from the radar equation and the values listed previously;  $S(F_s)$  is the amplitude of the normalized illumination pattern in the direction  $F_s$ ;  $\chi$  is a binary random variable that equals "0" or "1" when the target is absent or present, respectively; and  $\mathbf{n}$  is a length- $N$  complex Gaussian random vector with average power  $P_n$ . Under this model, the pdf of the measurements is known for both the target-absent and target-present cases. The pdf of the nuisance parameter  $F_s$  can be updated from measurements received according to

$$p(F_s|\mathbf{y}) = p(\mathbf{y}|F_s; \chi=1)p(F_s)/p(\mathbf{y}) \quad (3.39)$$

where the denominator in Eq. (3.39) is a normalization factor that does not need to be computed. Similarly, if we define the prior probability of target presence as  $P(\chi)$ , we can define posterior probabilities of target presence as

$$P(\chi|\mathbf{y}) = p(\mathbf{y}|\chi)P(\chi)/p(\mathbf{y})$$

where  $p(\mathbf{y}|\chi=1) = \int p(\mathbf{y}|F_s; \chi=1)p(F_s)dF_s$  and  $p(\mathbf{y}|\chi=0) = \mathcal{CN}(0, P_n \mathbf{I}_N)$ .

As in the matched illumination example of the previous sections, target detections are declared when the probability of target presence exceeds a prescribed threshold. In contrast to the previous example, however, multiple transmissions will be simulated and the nuisance parameter pdf and probability of target presence will be updated after each transmission.

If the target's location were known, the illumination pattern giving best detection results would be the one that is focused on the target with uniform amplitude and linear phase weights, i.e., matched to the array steering vector. However, this focused pattern sacrifices angular coverage, such that most directions are weakly illuminated. Here, we calculate an optimized illumination pattern on each transmission using the matched waveform theory described early in this chapter.

If the target is located at spatial frequency  $F_t$ , then a normalized spatial spectrum of the received signal power can be defined as  $|G(F_s, F_t)|^2 = \delta(F_s - F_t)$ . Using the pdf of the AOA, the expected spatial power spectrum is then

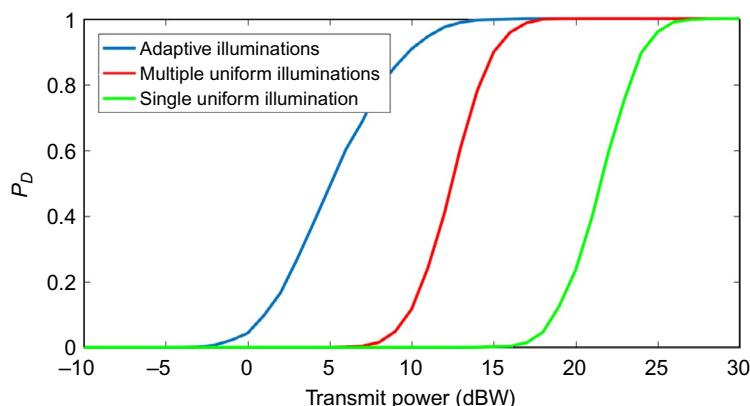
$$\Psi_{GG}(F_s) = \int |G(F_s, F_t)|^2 p(F_t) dF_t = \int \delta(F_s - F_t) p(F_t) dF_t = p(F_s). \quad (3.40)$$

This PSD can be substituted for the signal power spectrum in Eq. (3.10) in order to compute a kernel for use in the integral equation of Eq. (3.9), implemented in simulation as a discrete eigenvalue/eigenvector problem. The Fourier transform of the eigenvector associated with the largest eigenvalue is then taken as the normalized illumination pattern  $S(F)$ . On each illumination, the pdf of the target location is updated to reflect the most recent knowledge regarding target location. As this knowledge, quantified through  $p(F_s)$ , is updated, the transmit illumination pattern adapts in response, leading to a more effective search mode and better detection performance for the same transmit power.

### 3.5.2.2 Carryover and Adaptation Performance Gains

Fig. 3.8 shows detection performance results for the problem described previously with three different illumination strategies. The first strategy assumes a single illumination with uniformly distributed intensity across normalized spatial frequencies from  $-0.4$  to  $0.4$ . The second strategy uses eight transmissions, each with uniform illumination. The measurements produced on each illumination are used to update the probability of target presence, and the probability of target presence is compared to a threshold after the eighth illumination to decide whether to declare a target. In the final strategy, the first illumination is a uniform illumination over the span of possible target AOAs. However, subsequent illuminations are adapted to the expected spatial power spectrum. For the results in Fig. 3.8, the probability threshold needed to declare a detection was  $0.8$ .

In comparing the single illumination result to the result for multiple illuminations with a constant, uniformly distributed transmit pattern, we see that both performance curves have the same shape. The multiple illumination results are shifted to the left by  $9$  dB, which means that the multiple illumination strategy can achieve the same



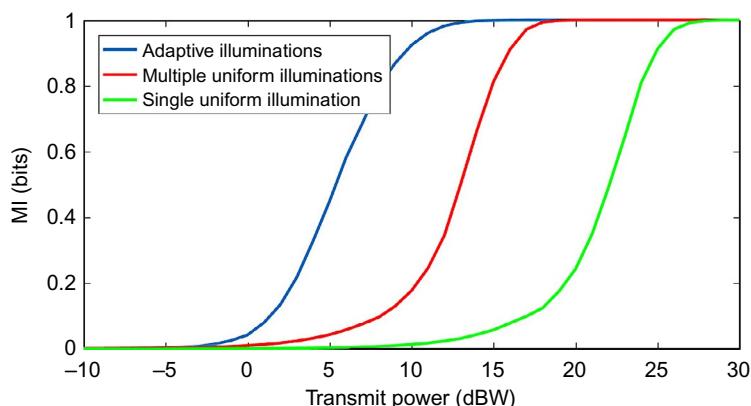
**FIG. 3.8**

Detection performance comparison for different transmit illumination strategies in a detection problem with unknown target direction.

detection performance as a single illumination, yet with 9 dB less transmit power per illumination. Of course, the  $10 \log_{10}(8) = 9$  dB shift is no coincidence as it corresponds to the exact amount of additional energy made available by having eight transmissions instead of one. Therefore the shift between the red and green curves in Fig. 3.8 shows the benefit of carrying information over between illuminations in the form of target probabilities. In a nondynamic, uniform illumination scenario, this benefit corresponds directly to the extra energy allocated to the detection problem. In a dynamic scenario, we would expect the gains to be lower due to losses associated with aging of information extracted from prior illuminations when the target parameters were different. The timescale and size of the scene dynamics compared to the measurement revisit rate will determine the “carryover gain” that can be achieved.

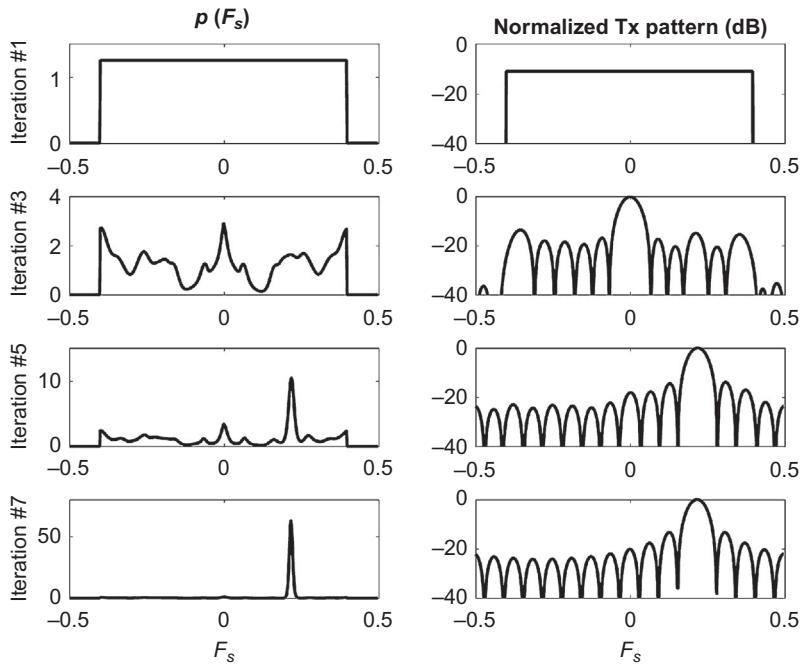
As seen in Fig. 3.8, closed-loop adaptive illumination provides additional benefit. The uniform, multiple-illumination strategy and the adaptive illumination strategy both carry information from one illumination to the next. Hence, they integrate information gained over multiple transmissions rather than a “single-shot” threshold comparison. Because they both carry information, the improved performance of the adaptive illumination is an “adaptation gain” produced by modifying the transmit pattern to illuminate the mostly likely target angles. As some angles or regions show promise for a potential target, subsequent illuminations focus more energy in those directions, thereby increasing SNR and improving the chances of target detection. For this scenario, the adaptation gain is nearly 10 dB at lower detection probabilities and a few dB at high detection probabilities (where even the spoiled reduced-gain transmit beam provides reasonable performance).

Fig. 3.9 shows corresponding results for MI gained on the detection problem, as quantified by the change in entropy of the binary random variable  $\chi$  over the course



**FIG. 3.9**

Detection-specific mutual information gained by the radar system versus transmit illumination power for the case of unknown target direction.

**FIG. 3.10**

Nuisance parameter (AOA) pdfs at various stages of an adaptive, multiillumination scenario. The transmit power was 5 dBW. The pdfs on the left yield corresponding transmit patterns on the right.

of the experiment. The results in Figs. 3.8 and 3.9 were obtained by averaging over 5000 Monte Carlo trials with independent generation of the target AOA on each trial.

Fig. 3.10 shows an example of how the nuisance parameter pdf changes over a sequence of measurements, along with the corresponding adaptation of the transmit illumination. The figures are taken from a single trial with 5-dBW transmit power per illumination, and the first, third, fifth, and seventh illuminations are shown. On the first illumination, the nuisance parameter (spatial frequency) pdf is uniformly distributed over the span of possible AOAs, and the transmit illumination pattern is an idealized uniform illumination over those same angles. On the third illumination, the prior two illuminations have provided information regarding the potential directions where a target might be located. Some directions have become less likely and others have become more likely. Substituting the updated pdf shown in the left panel into Eq. (3.40), the SNR-optimized transmit pattern is shown in the corresponding right panel. This optimized transmit pattern is found by substituting Eq. (3.40) for the power spectrum into the integral equation defined in Eqs. (3.8)–(3.10). It is clearly seen that the transmit pattern adapts in response to the likely target angles. The third illumination focuses to array broadside, which

is the most likely direction after the first two illuminations, but also has slightly raised sidelobes at other angles that are somewhat likely. After four transmissions, one narrow angular region has emerged as the most likely direction, and the final transmissions are focused in that direction. It is clear that the adaptive system focuses maximum gain on the target, when supported by the nuisance parameter pdf. This adaptive focusing enables the detection and information performance gains seen in Figs. 3.8 and 3.9. Currently, most traditional radar systems do not possess the capability for such adaptability, which requires advanced software, hardware, and computational capability. These sample results, however, indicate the potential gains achievable if future systems can be implemented with adaptive measurement paradigms.

### 3.5.3 PARALLEL ESTIMATION

The previous examples demonstrated benefits of cognitive radar through two fundamentally important detection problems. In these problems, adapting the radar's illumination parameters enhanced SNR, resulting in improved detection performance. However, radar systems must also estimate continuous variables such as radar target parameters or images of scattering. In many applications, the number of parameters to estimate or track may be large and must be performed with finite limitations on radar timeline and power. For example, the extensive proliferation of small, unmanned vehicles in public airspace may eventually overload the resources of radar systems that must track the vehicles for safety and security purposes. In this example, we demonstrate the potential of cognitive radar to improve tracking performance and behavior in a resource-constrained environment.

The canonical problem we demonstrate here is one where multiple, statistically independent parameters must be tracked over time. The parameters can only be updated when illuminated, and the uncertainty of the parameter value grows over time between updates. Thus although we will assume a simplified linear model, the abstracted problem considered here can be related to a multitarget tracking problem, and the language used will be in the context of estimating a single parameter of multiple different targets.

Let there be  $M$  parameters,  $\zeta_1 \dots \zeta_M$ , that must be estimated, one scalar parameter per target. The parameters are assumed to be Gaussian distributed with a variance of  $\sigma_{\zeta_m}^2$ , and we assume a linear model that relates the  $m$ th parameter to a measurement at the  $k$ th illumination interval. These assumptions model the situation where a radar operates in the asymptotic region of SNR (i.e., the SNR is higher than the threshold SNR present in nonlinear estimation problems) and parameter estimates are Gaussian distributed around the true parameter. Under this model, the measurement of the  $m$ th parameter on the  $k$ th illumination is

$$y_m[k] = \alpha_{m,k}\zeta_m + n_{m,k} \quad (3.41)$$

where  $\alpha_{m,k}$  depends on the strength of the  $k$ th illumination of the  $m$ th parameter and  $n_{m,k}$  is statistically independent, zero-mean Gaussian noise with variance  $\sigma_n^2$ .

In keeping with a radar example, we make the illumination strength related to power on target and the range to the target; therefore defining  $R_m$  as the range to the  $m$ th parameter, the illumination strength is

$$\alpha_{m,k} = \sqrt{\frac{P_m[k]}{R_m^4}} \quad (3.42)$$

where the illumination power  $P_m[k]$  will be varied according to various conventional and cognitive approaches in the simulations later. Obviously, the SNR of the parameter measurement increases with increased power on target and decreases with range to the target.

Ignoring the time index for the moment, the posterior variance of the parameter under the linear Gaussian model—that is, the variance of the estimate of  $\zeta_m$  given measurement of  $y_m$ —is given as [34]

$$\sigma_{\zeta_m|y_m}^2 = \sigma_{\zeta_m}^2 \left( \frac{\sigma_n^2}{\alpha_m^2 \sigma_{\zeta_m}^2 + \sigma_n^2} \right). \quad (3.43)$$

From Eq. (3.43), we can see that in the limit as the noise power goes to zero or as the illumination strength goes to infinity, the posterior variance goes to zero. This is the situation where infinite SNR produces a perfect parameter estimate. In contrast, if the noise power is large or if the target is not illuminated ( $\alpha_m = 0$ ), then the term in parenthesis goes to one and the posterior variance is the same as the prior variance. The information gained depends on the change in the variance because

$$\begin{aligned} I(y_m; \zeta_m) &= h(\zeta_m) - h(\zeta_m|y_m) \\ &= \frac{1}{2} \log 2\pi e \sigma_{\zeta_m}^2 - \frac{1}{2} \log 2\pi e \sigma_{\zeta_m|y_m}^2 \\ &= \frac{1}{2} \log \frac{\sigma_{\zeta_m}^2}{\sigma_{\zeta_m|y_m}^2} \end{aligned} \quad (3.44)$$

where  $h(\cdot)$  denotes entropy. Using Eq. (3.43), the MI gained is

$$I(y_m; \zeta_m) = \frac{1}{2} \log \left( \frac{\alpha_m^2 \sigma_{\zeta_m}^2 + \sigma_n^2}{\sigma_n^2} \right) = \frac{1}{2} \log \left( 1 + \frac{\alpha_m^2 \sigma_{\zeta_m}^2}{\sigma_n^2} \right). \quad (3.45)$$

Note that Eq. (3.45) is consistent with the form  $\frac{1}{2} \log(1 + SNR)$ , which is the expected form for linear observation of a Gaussian random variable in additive Gaussian noise. This is the same form that appears in the equation for the capacity of a Gaussian communication channel [35]. The difference here is that the variance of the variable to be estimated is separated from the illumination power, such that the information gained will depend on the additive noise power, the power that the radar places on the target, and the range from the radar to the target. In any radar application, the power available for allocation will be finite, meaning that it is not possible to achieve arbitrarily high SNR for each of the  $M$  variables that must be

estimated. A conventional radar system would likely choose to illuminate each target in succession with maximum power and gain, but in a resource-constrained scenario, the results later show that a smart, adaptive illumination strategy will outperform a fixed illumination strategy.

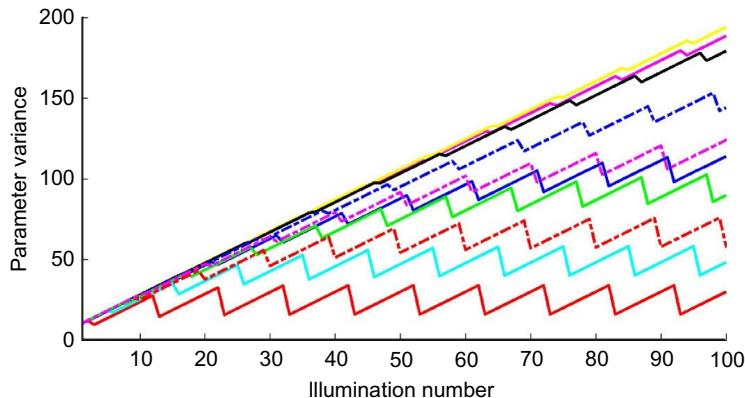
Returning to the time-indexed model, we note that Eq. (3.45) describes the information gained on any particular illumination, where  $\sigma_{\zeta_m}^2$  denotes the parameter variance just before illumination and  $\sigma_{\zeta_m|y_m}^2$  is the variance after illumination. In a static model, multiple observations would eventually drive the variance down to zero. However, we assume a simple dynamic model here where

$$\zeta_m[k] = \zeta_m[k - 1] + u_m \quad (3.46)$$

where  $u_m$  is the plant noise modeled as a zero-mean Gaussian random variable with variance  $\sigma_u^2$ . In other words, the variance of the parameter will increase by  $\sigma_u^2$  between illumination time steps. In essence, we have put together a dynamic model similar to the model used for Kalman tracking [34], but there is no higher-order motion to allow the target parameter's mean value to change over time. In the results later, we are only interested in how the parameter variance changes over time in response to different illumination strategies, not the actual parameter estimates.

The results in the following figures are based on a simulation with  $M = 10$  scalar parameters to be estimated. We assume that the targets corresponding to these parameters are located in unique spatial directions, such that power transmitted toward one target does not affect the others. If it is desired to illuminate more than one target in a single transmission, then the transmitted power budget must be shared between the targets (e.g., multiple beams). The 10 targets have randomly generated ranges from 1 to 10 km according to a uniform distribution. In the first results, the power budget is 20 dBW (although specific units are irrelevant here as we have absorbed other constants in the radar equation), the noise power is  $-100$  dBW, and the target variances are tracked over 100 illuminations. The initial variance of each target's parameter is 10 (arbitrary units). For every parameter, the process noise variance is  $\sigma_u^2 = 2$ .

Fig. 3.11 shows the variances for all 10 parameters over the 100 illuminations using a sequential illumination strategy. In the sequential strategy, all 20 dBW units are allocated toward the first target on illumination #1, toward the second target on illumination #2, and so on. After the  $M$ th target is illuminated, the process repeats again with the first target. The second target was the closest target, with a range of 2.3 km. The parameter variance for this target (solid red) varies between approximately 10 and 25. Between illuminations, the variance creeps up to its peak value before it is reduced again on every tenth illumination. The next closest target (corresponding to solid light blue) is located at 3.3 km and its variance has the same behavior but fluctuates around 50. The variance levels are higher for this target because of the lower SNR resulting from increased target range. Furthermore, there are three targets (yellow, pink, and black solid lines) at ranges of more than 8 km whose variance never stabilizes. The ranges to these targets are too far, such that one full-gain illumination out of every 10 is insufficient to combat the growth in

**FIG. 3.11**

Parameter variances over time for 10 independent scalar parameters and a sequential illumination strategy.

uncertainty from the process noise. In other words, illuminations of these targets while their variances are still relatively small are wasted—the measurement quality is insufficient to reduce the parameter variance by an appreciable amount. Even if every illumination was focused on one of these targets, it is not clear if the system has enough power and timeline to hold the parameter variance in check; therefore rather than wasting this transmitted power, a more efficient power allocation strategy should be used.

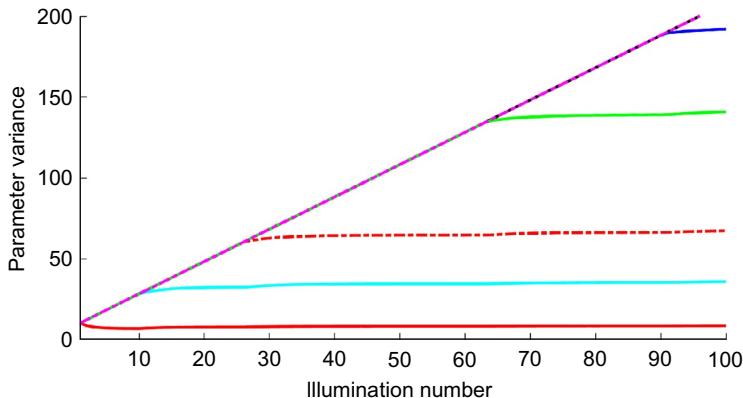
[Fig. 3.12](#) shows parameter variance over time using an illumination strategy that optimizes MI at each illumination. Using Eq. (3.45) and separating the illumination strength factors into the power and range contributions, the information gained on the  $m$ th parameter is

$$\frac{1}{2} \log \left( 1 + \frac{P_m \sigma_{\zeta_m}^2 / R_m^4}{\sigma_n^2} \right) = \frac{1}{2} \log \left( 1 + \frac{P_m}{\sigma_n^2 R_m^4 / \sigma_{\zeta_m}^2} \right). \quad (3.47)$$

It is then possible to define an information-optimization problem according to

$$\max_{P_1, P_2, \dots, P_M} \sum_{m=1}^M \frac{1}{2} \log \left( 1 + \frac{P_m}{\sigma_n^2 R_m^4 / \sigma_{\zeta_m}^2} \right) \text{ subject to } \sum_{m=1}^M |P_m|^2 = P_{tot}. \quad (3.48)$$

This is a well-known optimization problem for the capacity of parallel Gaussian channels [35]. Here, the differing range to each target and the separation of the parameter variance from the transmit power cause an effective variation in the SNR per channel. Therefore the formulation in Eq. (3.48) is similar to the capacity formulation for multiple-input, multiple-output (MIMO) communication channels

**FIG. 3.12**

Parameter variances over time for 10 independent scalar parameters and an information-optimal, adaptive illumination strategy.

where the individual channels have different channel gains. Using Lagrange multipliers, the resulting solution is

$$P_m = \left( A - \sigma_n^2 R_m^4 / \sigma_{\zeta_m}^2 \right)^+ \quad (3.49)$$

where  $A$  is found from the power constraint according to

$$\sum_{m=1}^M \left( A - \sigma_n^2 R_m^4 / \sigma_{\zeta_m}^2 \right)^+ = P_{tot} \quad (3.50)$$

where  $(\nu)^+$  is defined as

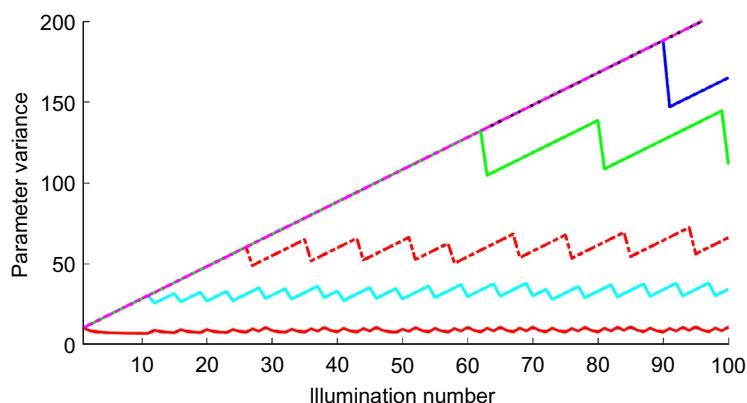
$$(\nu)^+ = \begin{cases} 0 & \nu < 0 \\ \nu & \nu \geq 0 \end{cases}$$

In other words, the individual channel illumination powers must be positive and the level  $A$  is chosen such that the power constraint is met.

In the information-optimal strategy, the parameter variances and target ranges are used at each illumination step to obtain an optimum distribution of power illuminating each target. On illumination, the variances are reduced according to the quality of the measurement on each parameter, and then the process noise increases the parameter variances prior to the next measurement. The results of such a strategy are shown in Fig. 3.12. The parameters, including individual target ranges, were the same as the ones used for Fig. 3.11. In Fig. 3.12 we see that the closest targets stabilize to a parameter variance that is smaller than in Fig. 3.11. For example, the closest target stabilizes to a value close to the initial variance of 10, and the second closest stabilizes at a variance of approximately 35. These values are lower than the oscillating variances for the same targets in the sequential illumination strategy. The trade-off is that long-range targets are not updated until their variance becomes very

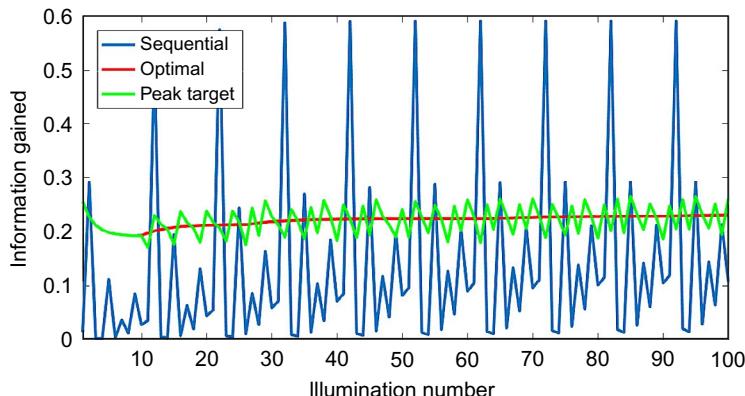
large. In contrast to Fig. 3.11, where multiple targets stabilize to a variance of around 100, only four targets stabilize to variances below 150 in Fig. 3.12. The reason is that from an information perspective, it is more efficient to tighten the parameter estimates on nearby, high-SNR targets than it is to spend valuable transmit power on targets that are too far away to keep the variance from growing. Note that in Fig. 3.11, the variances of the targets that appear to stabilize near a variance of 100 are actually still slowly growing. In the information-optimal approach, these targets are allowed to go unmeasured for long periods of time until their variances grow so large that even a low-SNR measurement is valuable. In fact, five of the targets never achieve, within the 100 illuminations shown, a variance that is large enough to be worthy of an update. Remember that because range factors into radar SNR according to its fourth power, a target that is 10 times further away than an equivalent target will have 40 dB lower SNR.

In many cases, a radar system may not be capable of illuminating multiple targets with an arbitrary fraction of the available transmit power. Thus the adaptive procedure was modified in Fig. 3.13 to be based on the individual target that would provide the most MI at each transmit interval. Prior to each illumination, the expected information gained from each target is calculated, and the transmit power is allocated to the target that is expected to produce the most information. Notice in Fig. 3.13 that only the closest target is illuminated over approximately the first 10 illuminations while the variance of the other nine targets is allowed to grow. However, once the second-closest target reaches a critical variance, the radar system begins alternating between those targets. Eventually, a third, fourth, and fifth target enter into the alternating illumination but at different update rates. The targets that are furthest away provide the least information; therefore they are illuminated less often. Although their absolute drops in variance per illumination are larger (e.g., compare



**FIG. 3.13**

Parameter variances over time for 10 independent scalar parameters and a peak information illumination strategy (single target that will produce the most information).

**FIG. 3.14**

Information gained per illumination for each of the three illumination strategies.

the solid green and light blue curves), the relative changes are approximately equal, resulting in approximately equal information gained.

Finally, Fig. 3.14 shows a comparison of the MI obtained on each illumination for each of the illumination strategies. The information obtained via the sequential target strategy fluctuates considerably as the radar illuminates targets at different ranges. The information gained by the optimal strategy converges to a steady level determined by the specific target ranges and power available, while the peak information strategy has small fluctuations around the optimal level. In summing the information gained over the experiment, the optimal and peak strategies are nearly equal and collect approximately 50% more information than the sequential strategy.

### 3.5.4 SUMMARY

The basic simulation experiments performed in this section are meant to demonstrate the trade-offs and potential benefits of customizing measurements in an adaptive, closed-loop fashion. Given that any sensing problem will face resource constraints, these types of optimization strategies can make radar systems more efficient in their use of available power, timeline, and other degrees of freedom. For example, in the “tracking” problem, it was more information efficient to maintain tighter track on a few closely located targets than to try to maintain track on all targets. Although it may seem unsatisfying to watch track variances grow un-illuminated, the reality of a target-dense resource-constrained environment is that trying to accomplish everything may lead to failure on everything. Furthermore, one could easily envision applying additional constraints to the optimization problem, such as specifying sufficient track qualities (i.e., minimum variances) beyond which no updates are needed, or worst-case track qualities that should be given extra consideration to prevent track loss.

Many different applications and approaches are possible. The consistent features, however, must include methods to quantify desirable knowledge, current state of the propagation environment, models that predict expected outcomes over certain illumination parameters, and measurement optimization metrics.

---

### 3.6 COGNITIVE RADAR EXPERIMENTS

Cognitive radar experiments are exceptionally difficult to perform due to the adaptive nature of the sensing involved. In contrast to development of algorithms to be applied to already-captured data, it can be difficult to obtain appropriate data sets to prove cognitive radar concepts. In algorithm development, one can derive or conceive of improved processing steps, perform initial validation of the processing via simulation or other methods, and then apply the new processing techniques to standard experimental data sets. With cognitive radar, however, the measurement design is not static. Rather, the measurements are adapted in the loop, such that current measurements will change the course of future ones. The only ways to have this ability are (1) to fully implement a cognitive radar (though not necessarily capable of operating in real time), complete with signal processing, representation mechanism, feedback to the transmitter, and transmitter agility for some type of optimization, or (2) to capture an overdetermined data set that will be downselected in postcollection processing according to the cognitive techniques—in other words, to mimic a cognitive data collection after the fact.

Thus far, the main result in the literature for real data experiments of cognitive radar is contained in Ref. [50]. In this paper, the authors report on a test bed called the “Cognitive Radar Engineering Workspace” (CREW), which was used to demonstrate real-time adaptation of a radar system’s pulse repetition frequency (PRF) and CPI in order to meet desired tracking error metrics. The PRF was adjusted to avoid Doppler aliasing and mainlobe clutter, whereas the CPI was adapted to meet SNR constraints. For target tracking methods that maintain an estimate of the target SNR, this type of adaptation can be used to minimize the radar timeline dedicated to a single target while still achieving tracking goals. The time saved could then be applied toward other tasks.

---

### REFERENCES

- [1] S. Haykin, Cognitive radar: a way of the future, *IEEE Signal Process. Mag.* 23 (1) (2006) 30–40.
- [2] P. Stinco, M.S. Greco, F. Gini, Spectrum sensing and sharing for cognitive radars, *IET Radar Sonar Navig.* 10 (3) (2016) 595–602.
- [3] M.S. Greco, F. Gini, P. Stinco, Cognitive radars: some applications, in: *Proceedings 2016 IEEE Global Conference on Signal and Information Processing*, Washington, DC, December 7–9, 2016, pp. 1077–1082.

- [4] S. Alsaif, G.E. Smith, C.J. Baker, Using cognitive radar to traverse apertures, in: Proceedings 2014 International Radar Conference, Lille, October 13–17, 2014, pp. 1–6.
- [5] D.R. Fuhrmann, Active-testing surveillance systems, or, playing twenty questions with a radar, in: Proceedings 11th Annual Adaptive Sensor and Array Processing Workshop, Lexington, March 11–13, 2003.
- [6] N.A. Goodman, P.R. Venkata, M.A. Neifeld, Adaptive waveform design and sequential hypothesis testing for target recognition with active sensors, *IEEE J. Sel. Top. Sign. Process.* 1 (1) (2007) 105–113.
- [7] K.L. Bell, C.J. Baker, G.E. Smith, J.T. Johnson, M. Rangaswamy, Cognitive radar framework for target detection and tracking, *IEEE J. Sel. Top. Sign. Process.* 9 (8) (2015) 1427–1439.
- [8] S. Haykin, Cognitive radar networks, in: Proceedings 1st IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing, Puerto Vallarta, December 13–15, 2005, pp. 1–3.
- [9] D.R. Fuhrmann, L. Boggio, Active-testing surveillance for multiple target detection with composite hypotheses, in: Proceedings 2003 IEEE Workshop on Statistical Signal Processing, St. Louis, September 2003, pp. 641–644.
- [10] M.R. Bell, Information theory and radar waveform design, *IEEE Trans. Inf. Theory* 39 (5) (1993) 1578–1597.
- [11] S.U. Pillai, H.S. Oh, D.C. Youla, J.R. Guerci, Optimum transmit-receiver design in the presence of signal-dependent interference and channel noise, *IEEE Trans. Inf. Theory* 46 (2) (2000) 577–584.
- [12] D.A. Garren, M.K. Osborn, A.C. Odom, J.S. Goldstein, S.U. Pillai, J.R. Guerci, Enhanced target detection and identification via optimized radar transmission pulse shape, *Proc. IEEE* 148 (3) (2001) 130–138.
- [13] E. Mosca, Probing signal design for linear channel identification, *IEEE Trans. Inf. Theory* 18 (4) (1972) 481–487.
- [14] A. Wald, *Sequential Analysis*, Wiley, New York, 1947.
- [15] A. Wald, Sequential tests of statistical hypotheses, *Ann. Math. Stat.* 16 (2) (1945) 117–186.
- [16] P. Armitage, Sequential analysis with more than two alternative hypotheses, and its relation to discriminant function analysis, *J. R. Stat. Soc. Ser. B* 12 (1) (1950) 137–144.
- [17] C.W. Baum, V.V. Veeravalli, A sequential procedure for multihypothesis testing, *IEEE Trans. Inf. Theory* 40 (6) (1994) 1994–2007.
- [18] V.P. Dragalin, A.G. Tartakovsky, V. Veeravalli, Multihypothesis sequential probability ratio tests – part I: asymptotic optimality, *IEEE Trans. Inf. Theory* 45 (7) (1999) 2448–2461.
- [19] V.P. Dragalin, A.G. Tartakovsky, V. Veeravalli, Multihypothesis sequential probability ratio tests – part II: accurate asymptotic expansions for the expected sample size, *IEEE Trans. Inf. Theory* 46 (4) (1999) 1366–1383.
- [20] A.G. Tartakovsky, Asymptotic optimality of certain multihypothesis sequential tests: non-i.i.d. case, *Stat. Infer. Stoch. Process.* 1 (3) (1998) 265–295.
- [21] J.A. Simmons, The resolution of target range by echolocating bats, *J. Acoust. Soc. Am.* 54 (1) (1973) 157–173.
- [22] J.A. Thomas, C.F. Moss, M. Vater, *Echolocation in Bats and Dolphins*, University Chicago Press, Chicago, 2004.

- [23] C.J. Baker, M. Vespe, G.I. Jones, Target classification by echo locating animals, in: Proceedings 2007 International Waveform Diversity and Design Conference, Pisa, June 4–8, 2007, pp. 348–352.
- [24] M. Hurtado, A. Nehorai, Bat-inspired adaptive design of waveform and trajectory for radar, in: Proceedings 2008 42nd Asilomar Conference on Signals, Systems and Computers, Pacific Grove, October 26–29, 2008, pp. 36–40.
- [25] M. Vespe, G. Jones, C.J. Baker, Lessons for radar, *IEEE Signal Process. Mag.* 26 (1) (2009) 65–75.
- [26] A. Balleri, H. Griffiths, M. Holderied, C. Baker, Bat-inspired multi-harmonic waveforms, in: Proceedings 2010 International Waveform Diversity and Design Conference, Niagara Falls, August 8–13, 2010, pp. 86–89.
- [27] J.R. Guerci, *Cognitive Radar: The Knowledge-Aided Fully Adaptive Approach*, Artech House, Norwood, MA, 2010.
- [28] W.L. Melvin, G.A. Showman, J.R. Guerci, A knowledge-aided GMTI detection architecture, in: Proceedings 2004 IEEE Radar Conference, Philadelphia, April 26–29, 2004, pp. 301–306.
- [29] W.L. Melvin, J.R. Guerci, Knowledge-aided signal processing: a new paradigm for radar and other advanced sensors, *IEEE Trans. Aerosp. Electron. Syst.* 42 (3) (2006) 983–996.
- [30] A. Leon-Garcia, *Probability and Random Processes for Electrical Engineering*, Addison-Wesley, Reading, MA, 2004.
- [31] J.R. Guerci, *Optimal Radar Waveform Design*, vol. 2, Academic Press Library in Signal Processing, 2014, pp. 729–758.
- [32] R.A. Romero, J. Bae, N.A. Goodman, Theory and application of SNR and mutual information matched illumination waveforms, *IEEE Trans. Aerosp. Electron. Syst.* 47 (2) (2011) 912–927.
- [33] P.Z. Peebles, *Radar Principles*, John Wiley & Sons, New York, 1998.
- [34] S.M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*, Prentice-Hall, Upper Saddle River, NJ, 1993.
- [35] T.M. Cover, J.A. Thomas, *Elements of Information Theory*, second ed., Wiley, Hoboken, NJ, 2006.
- [36] R.A. Romero, N.A. Goodman, Cognitive radar network: cooperative adaptive beamsteering for integrated search-and-track application, *IEEE Trans. Aerosp. Electron. Syst.* 49 (2) (2013) 915–931.
- [37] S.U. Pillai, K.Y. Li, H. Beyer, Construction of constant envelope signals with given Fourier transform magnitude, in: Proceedings 2009 IEEE Radar Conference, Pasadena, May 4–8, 2009, pp. 1–4.
- [38] S.U. Pillai, K.Y. Li, B. Himed, Constant envelope signals with prescribed discrete Fourier transform magnitude, in: Proceedings 2011 IEEE Radar Conference, Kansas City, 23–27 May, 2011, pp. 470–473.
- [39] L.K. Patton, B.D. Rigling, Modulus constraints in adaptive radar waveform design, in: Proceedings 2008 IEEE Radar Conference, Rome, May 26–30, 2008, pp. 1–6.
- [40] E. Chong, C. Kreucher, A.O. Hero, Partially observable Markov decision process approximations for adaptive sensing, *Discret. Event Dyn. Syst.* 19 (2009) 377–422.
- [41] A. Charlish, F. Hoffmann, Anticipation in cognitive radar using stochastic control, in: Proceedings 2015 IEEE Radar Conference, Arlington, May 10–15, 2015, pp. 1692–1697.

- [42] B.L. Scala, B. Moran, R. Evans, Optimal adaptive waveform selection for target detection, in: Proceedings 2003 International Radar Conference, Adelaide, September 3–5, 2003, pp. 492–496.
- [43] B.L. Scala, M. Rezaeian, B. Moran, Optimal adaptive waveform selection for target tracking, in: Proceedings 8th International Conference on Information Fusion, Philadelphia, July 25–28, 2005, pp. 552–557.
- [44] S.M. Kay, Fundamentals of Statistical Signal Processing: Detection Theory, Prentice-Hall, Upper Saddle River, NJ, 1998.
- [45] M.A. Neifeld, A. Ashoka, P.K. Baheti, Task-specific information for imaging system analysis, *J. Opt. Soc. Am. A* 24 (12) (2007) B25–B41.
- [46] A. Ashok, P.K. Baheti, M.A. Neifeld, Compressive imaging system design using task-specific information, *Appl. Optics* 47 (25) (2008) 4457–4471.
- [47] Y. Boers, J.N. Driessens, Particle filter based detection for tracking, in: Proceedings American Control Conference, Arlington, June 25–27, 2001, pp. 4393–4397.
- [48] M.G. Rutten, N.J. Gordon, S. Maskell, Recursive track-before-detect with target amplitude fluctuations, *IEE Proc. Radar Sonar Navig.* 152 (5) (2005) 345–352.
- [49] M.G. Rutten, B. Ristic, N.J. Gordon, A comparison of particle filters for recursive track-before-detect, in: Proceedings 2005 8th International Conference on Information Fusion, Philadelphia, July 25–28, 2005, pp. 169–175.
- [50] G.E. Smith, Z. Cammenga, A. Mitchell, K.L. Bell, M. Rangaswamy, J.T. Johnson, C. J. Baker, Experiments with cognitive radar, in: Proceedings 2015 IEEE 6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing, Cancun, December 13–16, 2015, pp. 293–296.

# Parameter bounds under misspecified models for adaptive radar detection

# 4

**Stefano Fortunati, Fulvio Gini, Maria S. Greco**

*University of Pisa, Pisa, Italy*

## 4.1 LIST OF SYMBOLS AND FUNCTIONS

- $\mathbf{x} = \{\mathbf{x}_m\}_{m=1}^M$ : set of the available observation vectors assumed to be independent and identically distributed (iid).
- $p_X(\mathbf{x}) \triangleq p_X(\mathbf{x}; \boldsymbol{\tau})$ : true joint probability density function (pdf) possibly parameterized by a real (deterministic and unknown) *true* vector  $\boldsymbol{\tau} \in \mathcal{T} \subset \mathbb{R}^p$ .
- $f_X(\mathbf{x}) \triangleq f_{\boldsymbol{\theta}}(\mathbf{x}) \triangleq f_X(\mathbf{x}; \boldsymbol{\theta})$ : assumed joint pdf parameterized by a real (deterministic and unknown) vector  $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^d$ .
- $\hat{\boldsymbol{\theta}}(\mathbf{x})$ : estimator of the deterministic parameter vector  $\boldsymbol{\theta}$  based on the data  $\mathbf{x}$ .
- $\nabla_{\boldsymbol{\theta}} u(\boldsymbol{\theta}) \triangleq (\partial u / \partial \theta_1 \ \dots \ \partial u / \partial \theta_d)^T$ : gradient (column) vector of the scalar function  $u$ . With the notation  $\nabla_{\boldsymbol{\theta}_0} u(\boldsymbol{\theta}_0)$ , we define the gradient of the function  $u$  evaluated at  $\boldsymbol{\theta}_0$ .
- $\mathbf{U}_{\boldsymbol{\theta}} = \nabla_{\boldsymbol{\theta}}^T \mathbf{u}(\boldsymbol{\theta})$ : Jacobian matrix of the vector function  $\mathbf{u}$ . With the notation  $\mathbf{U}_{\boldsymbol{\theta}_0}$ , we define the Jacobian matrix of the function  $\mathbf{u}$  evaluated at  $\boldsymbol{\theta}_0$ .
- $E_p\{\mathbf{u}\} = \int \mathbf{u}(\mathbf{x}) p_X(\mathbf{x}) d\mathbf{x}$ : Expectation operator of the (scalar or vector) function  $\mathbf{u}$  with respect to a pdf  $p_X(\mathbf{x})$ .
- $D(p_X \| f_X) = \int p_X(\mathbf{x}) \ln \left( \frac{p_X(\mathbf{x})}{f_X(\mathbf{x})} \right) d\mathbf{x}$ : Kullback-Leibler divergence between  $p_X(\cdot)$  and  $f_X(\cdot)$ .
- $\text{vec}(\mathbf{A})$ : The vec-operator transforms a  $N \times N$  matrix  $\mathbf{A}$  into a vector by stacking the columns of the matrix one underneath the other.
- $\text{vecs}(\mathbf{A})$ : The vecs-operator denotes the  $N(N+1)/2 \times 1$  vector that is obtained from  $\text{vec}(\mathbf{A})$  by eliminating all supradiagonal elements of  $\mathbf{A}$ . For notation simplicity,  $N(N+1)/2 \triangleq l$ .
- $\mathbf{D}_N$ : Duplication matrix of order  $N$ . The duplication matrix is implicitly defined as the unique  $N^2 \times l$  matrix that satisfies the following equality  

$$\mathbf{D}_N \text{vecs}(\mathbf{A}) = \text{vec}(\mathbf{A}) \quad \text{for any symmetric matrix } \mathbf{A}.$$
- $\otimes$ : Kronecker product.
- $\times$ : Cartesian product.
- $\mathbf{A}^\dagger = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T$  is the Moore-Penrose pseudo-inverse of a matrix  $\mathbf{A}$ .

## 4.2 INTRODUCTION

The problem of estimating a deterministic parameter vector from a set of acquired data is ubiquitous in signal processing applications. A fundamental assumption underlying most estimation problems is that the true data model and the model adopted to derive an estimation algorithm are the same, that is, the model is correctly specified. However, a certain amount of mismatch is often inevitable in practice. Among others, the model mismatch can be due to an imperfect knowledge of the true data model or to the need to fulfill some operative constraints on the estimation algorithm (processing time, simple hardware implementation, etc.).

The first fundamental result on the general theory of the estimation under misspecification was provided by Huber in his seminal paper [1] on the statistical analysis of the maximum likelihood (ML) estimator under mismatched condition. This work was further developed by White in Ref. [2,3], where the term “quasi maximum likelihood” (QML) estimator was introduced. In particular, Huber and White have shown that the asymptotic distribution of the ML estimator under misspecified models is a Gaussian distribution. Further, the mean value of the ML estimator is the minimizer (also called *pseudo-true parameter vector* in Ref. [1]) of the Kullback-Leibler (KL) divergence between the true and the assumed data distributions, whereas the covariance matrix is given by the so-called Huber “sandwich” matrix. For the sake of clarity, in the rest of this chapter, we refer to the ML estimator under mismatched conditions as the mismatched maximum likelihood (MML) estimator. A milestone on the theoretical misspecification analysis is the book in Ref. [4] that covers all developments in this field from both the statistical and econometrical points of view. This book provides an excellent and insightful discussion about statistical inference in the presence of distributional misspecification, with a focus on estimation and hypotheses testing problems.

In conjunction with the asymptotic analysis of the ML estimator, a question that naturally arises is whether it is possible to establish a lower bound on the error covariance matrix of a certain class of mismatched estimators. When the true parametric model is specified, few of such lower bounds exist; one of these is the well-known Cramér-Rao bound (CRB). In a pioneering working paper [5], Q. H. Vuong first proposed a generalization of the CRB to the estimation problem under misspecified models.

Quite surprisingly and despite of the wide variety of potential applications, these fundamental results, disseminated in the statistical and econometrical literature, have remained largely unrecognized by the Signal Processing Community for many years. Only recently, the findings about the estimation theory under misspecification have been rediscovered and applied to different signal processing problems. In a recent book [6], the main results on the ML estimator under misspecified models have been reported. A different mismodeling related to the dynamic of the acquired data has been investigated in Ref. [7]. In particular, the asymptotic performance of the ML estimator and of the generalized likelihood ratio test (GLRT) is derived under the assumption of

independent identically distribution (iid) samples, when these samples are correlated in the actual model. Other recent works attempt to generalize the Cramér-Rao inequality in the presence of model misspecification. In Ref. [8], a Bayesian bound of the Ziv-Zakai type has been derived under model mismatch conditions restricted to misparameterized zero mean complex Gaussian distributions. On the same line, in Refs. [9–11], a Ziv-Zakai bound in the presence of model misspecification has been discussed and applied to the time-of-arrival (TOA) estimation problem. Recently, Richmond and Horowitz (first in Ref. [12] and then in Ref. [13]) derived a covariance inequality for deterministic complex parameter vector in the presence of model misspecification and introduced the term misspecified Cramér-Rao bound (MCRB). Moreover, in Ref. [13] and in Ref. [14], a generalization to the mismatched case of the Slepian and Bangs formulae for the evaluation of the bound for multivariate complex Gaussian distributed observations is also derived (see Appendices A and B at the end of this chapter) and applied to the classical direction of arrival (DOA) estimation problem. Furthermore, the extension of the Slepian-Bangs (SB) formulae and of the related MCRB to the non-Gaussian setting is addressed in [75] where SB formulae for Complex Elliptically Symmetric (CES) distributions under model misspecification are derived. The application of the MCRB to the DOA estimation problem is also discussed in Ref. [15]. To the best of our knowledge, Ref. [13] represents the first attempt to introduce an organic framework for deriving a covariance inequality of the Cramér-Rao type in the presence of model mismatch to the Signal Processing Community. More recently, Richmond and Basu have extended the work in Ref. [13] to the Bayesian estimation framework [16,17]. Finally, the recent paper [18] is pertinent to misspecified bounds, where the authors adopted a different definition of unbiasedness and a different score function than in Refs. [5,13].

The aim of this chapter is twofold: in the first part, we provide a comprehensive review of the main findings about the MCRB and the MML estimator for deterministic parameter estimation. Two toy examples are also provided in order to clarify the main theoretical concepts. In the second part, we discuss the application of the MCRB and of the MML estimator to a practical radar signal processing problem: the estimation of the disturbance covariance (scatter) matrix for adaptive radar detection [19,20]. We recast this classical radar problem in the more general context of the estimation of the scatter matrix in the complex elliptically symmetric (CES) distribution family [21].

The rest of the chapter is organized as follows. Section 4.3 provides the formal description of the general deterministic estimation problem under model misspecification. In Section 4.4, the main theoretical results on the MCRB and the MML estimator are reviewed and discussed. In Section 4.5, two simple toy examples are described to better clarify the theoretical findings of Section 4.4 and how they should be applied. Section 4.6 focuses on the application of the MML estimator and of the MCRB to the estimation of the scatter matrix in the CES distribution family, while Section 4.7 discusses the application of the MML scatter matrix estimator in adaptive radar detection problems. Section 4.8 summarizes our conclusions.

---

### 4.3 PROBLEM STATEMENT AND MOTIVATIONS

In the following, a formal description of the estimation problem under mismatched conditions is provided. Let  $\mathbf{x}_m \in \mathbb{C}^N$  be a  $N$ -dimensional random vector representing the outcome of a random experiment (i.e., the observation vector) with cumulative distribution function (cdf)  $P_X(\mathbf{x}_m)$ . In the remainder of the chapter, we assume that  $P_X(\mathbf{x}_m)$  has a relevant probability density function (pdf)  $p_X(\mathbf{x}_m)$ , and we use, with a small abuse of definition, the term “distribution” always to indicate the relevant pdf. Assume that the true pdf of  $\mathbf{x}_m$  is known to belong to a family  $\mathcal{P}$ . A *structure*  $T$  is a set of hypotheses, which implies a unique pdf in  $\mathcal{P}$  for  $\mathbf{x}_m$ . Such pdf is indicated with  $p_X(\mathbf{x}_m; T)$  [22,23]. The set of all the a priori possible structures for  $p_X$  is called a *model*. We assume that the pdf of the random vector  $\mathbf{x}_m$  has a parametric representation, i.e., we assume that every structure  $T$  is parameterized by a  $d$ -dimensional vector  $\boldsymbol{\tau}$  and that the model is described by a compact subspace  $T \subset \mathbb{R}^d$ .

The common assumption underlying any practical estimation problem is the perfect knowledge of the (joint) pdf  $p_X(\mathbf{x}; \boldsymbol{\tau})$  that characterizes the iid observations,  $\mathbf{x} = \{\mathbf{x}_m\}_{m=1}^M$ , except for the value of the parameter vector  $\boldsymbol{\tau} \in T$ . However, a certain amount of mismatch between the true pdf of the observations and the pdf assumed to derive an estimator of the parameters of interest is always present. Specifically, suppose that the true parametric pdf of the observations,  $p_X(\mathbf{x}; \boldsymbol{\tau})$ , and the assumed pdf,  $f_X(\mathbf{x}; \boldsymbol{\Theta})$ , with  $\boldsymbol{\Theta} \in \Theta \subset \mathbb{R}^d$ , belong to two (generally different) families of pdf's,  $\mathcal{P}$  and  $\mathcal{F}$ , that are parameterized by two possibly different parameters spaces  $T$  and  $\Theta$ :

$$\mathcal{P} = \{p_X | p_X(\mathbf{x}; \boldsymbol{\tau}) \text{ is a pdf } \forall \boldsymbol{\tau} \in T\}, \quad \mathcal{F} = \{f_X | f_X(\mathbf{x}; \boldsymbol{\Theta}) \text{ is a pdf } \forall \boldsymbol{\Theta} \in \Theta\}.$$

It is worth noting that the true parameter space  $T$  and the assumed parameter space  $\Theta$  may be completely different and/or have a different dimensions.

Since, in practical situations, the true model is unknown, i.e., we have no prior information on the particular parameterization of the true distribution, in the following, we refer to  $p_X(\mathbf{x}; \boldsymbol{\tau})$  only as  $p_X(\mathbf{x})$  in order to highlight the fact the neither the model, nor the true parameter vector  $\boldsymbol{\tau}$  is accessible by a mismatched estimator [13,20].

Suppose then that the  $M$  (possibly complex) iid measurement vectors are sampled from a particular pdf belonging to  $\mathcal{P}$ , i.e.,  $\mathbf{x}_m \sim p_X(\mathbf{x}_m)$ , for  $m = 1, 2, \dots, M$ . Due to a lack of knowledge or to the need to fulfill some computational requirements, a parametric pdf  $f_X(\mathbf{x}; \boldsymbol{\Theta})$ , belonging to the family  $\mathcal{F}$ , is assumed for the dataset  $\mathbf{x}$ . In this case a possible inference algorithm, e.g., an estimation algorithm, may be based on a *misspecified data model*, i.e., on the assumed pdf  $f_X(\mathbf{x}; \boldsymbol{\Theta})$  and not on the true pdf  $p_X(\mathbf{x})$ . The question that arises is to how will the statistical properties of an estimator (e.g., convergence, consistency, efficiency) defined in the classical estimation framework change in this mismatched scenario? This is the main topic of the next section.

---

## 4.4 A GENERALIZATION OF THE DETERMINISTIC ESTIMATION THEORY UNDER MODEL MISSPECIFICATION

The aim of this section is to provide an organic view of the findings in Refs. [1–3,5,13,20]. Starting from Ref. [5], we first provide a list of regularity conditions that are not only a fundamental prerequisite for the derivation of the MCRB but also allow better understanding of the nature and the usefulness of this bound. Then, we provide the expression of the MCRB and the class of estimators to which it applies (Theorem 4.1 [5]). Finally, we conclude this section by introducing the MML estimator, its asymptotic properties, and their link with the MCRB.

### 4.4.1 REGULAR MODELS

As stated before, let  $\mathbf{x} = \{\mathbf{x}_m\}_{m=1}^M$  be a set of iid  $N$ -dimensional random vectors and let  $p_X(\mathbf{x})$  the true pdf of  $\mathbf{x}$ . Let  $\mathcal{F} = \{f_X(\mathbf{x}; \boldsymbol{\theta}) \text{ is a p.d.f. } \forall \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d\}$  be a family of parametric pdfs that possibly does not contain  $p_X(\mathbf{x})$ .

**Assumption 4.1** For every  $\boldsymbol{\theta} \in \Theta$ , the functions  $|\ln f_X(\mathbf{x}; \boldsymbol{\theta})|$ ,  $|\partial \ln f_X(\mathbf{x}; \boldsymbol{\theta})/\partial \theta_i|$ , and  $|\partial^2 \ln f_X(\mathbf{x}; \boldsymbol{\theta})/\partial \theta_i \partial \theta_j|$ ,  $i, j = 1, \dots, p$ , are dominated by a function  $m(\mathbf{x})$  independent of  $\boldsymbol{\theta}$  and square-integrable with respect to  $p_X(\mathbf{x})$ .

**Assumption 4.2** (a) The function  $\zeta(\boldsymbol{\theta}) \triangleq E_p \{ \ln f_X(\mathbf{x}_m; \boldsymbol{\theta}) \} = \int \ln f_X(\mathbf{x}_m; \boldsymbol{\theta}) p_X(\mathbf{x}_m) d\mathbf{x}_m$  has a unique maximum on  $\Theta$  at an interior point  $\boldsymbol{\theta}_0$ . (b) The matrix  $\mathbf{A}_{\boldsymbol{\theta}_0}$  whose entries are

$$[\mathbf{A}_{\boldsymbol{\theta}_0}]_{ij} \triangleq \left[ E_p \left\{ \nabla_{\boldsymbol{\theta}_0} \nabla_{\boldsymbol{\theta}_0}^T \ln f_X(\mathbf{x}_m; \boldsymbol{\theta}_0) \right\} \right]_{ij} = E_p \left\{ \frac{\partial^2}{\partial \theta_i \partial \theta_j} \ln f_X(\mathbf{x}_m; \boldsymbol{\theta}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \right\} \quad (4.1)$$

is nonsingular. Note that  $\nabla_{\boldsymbol{\theta}_0} u(\boldsymbol{\theta}_0)$  indicates the gradient (column) vector of the scalar function  $u$  evaluated in  $\boldsymbol{\theta}_0$ . This can be recognized also as the identifiability condition (see [22–25]) for  $\boldsymbol{\theta}_0$ . The interior point  $\boldsymbol{\theta}_0$  can be equivalently seen as the point that minimizes the Kullback-Leibler divergence between the true distribution  $p_X(\mathbf{x}_m)$  and the assumed distribution  $f_X(\mathbf{x}_m; \boldsymbol{\theta})$  [3,4]:

$$\boldsymbol{\theta}_0 = \arg \min_{\boldsymbol{\theta} \in \Theta} \{D(p_X \| f_{\boldsymbol{\theta}})\} = \arg \min_{\boldsymbol{\theta} \in \Theta} \{-E_p \{ \ln f_X(\mathbf{x}_m; \boldsymbol{\theta}) \}\}, \quad (4.2)$$

where

$$D(p_X \| f_{\boldsymbol{\theta}}) \triangleq E_p \left\{ \ln \left( \frac{p_X(\mathbf{x}_m)}{f_X(\mathbf{x}_m; \boldsymbol{\theta})} \right) \right\} = \int \ln \left( \frac{p_X(\mathbf{x}_m)}{f_X(\mathbf{x}_m; \boldsymbol{\theta})} \right) p_X(\mathbf{x}_m) d\mathbf{x}_m. \quad (4.3)$$

**Assumption 4.3** There exists a neighborhood  $\Gamma$  of  $\boldsymbol{\theta}_0$  such that for every  $\boldsymbol{\theta} \in \Gamma$  the functions  $(f_X(\mathbf{x}; \boldsymbol{\theta}_0))^{-1} |\partial \ln f_X(\mathbf{x}; \boldsymbol{\theta})/\partial \theta_i|$ ,  $i = 1, \dots, p$  are dominated by a function  $m(\mathbf{x})$  independent of  $\boldsymbol{\theta}$  and square-integrable with respect to  $p_X(\mathbf{x})$ .

**Assumptions 4.1** and **4.3** essentially allow differentiation under the integral sign of the expectation of any random variable or vector with finite variance.

[Assumption 4.2](#) ensures the existence and the uniqueness of the so-called *pseudo-true parameters vector*  $\boldsymbol{\theta}_0$ . As seen later in the chapter,  $\boldsymbol{\theta}_0$  plays a key role both in the definition of the MCRB and of the MML.

**Definition 4.1 Regular models** [5] A parametric model  $\mathcal{F}$  is regular with respect to (w.r.t.) the pdf  $p_X(\mathbf{x})$  if [Assumptions 4.1–4.3](#) hold. It is regular w.r.t. a family  $\mathcal{P}$  if it is regular w.r.t. every pdf in  $\mathcal{P}$ . It is referred as regular if the regularity is w.r.t. every pdf in  $\mathcal{F}$ .

The following lemma summarizes some useful properties of parametric models that are regular w.r.t. the pdf  $p_X(\mathbf{x})$ . For any  $p_X(\mathbf{x})$  in  $\mathcal{P}$ , we define the matrix  $\mathbf{B}_{\boldsymbol{\theta}}$  as:

$$[\mathbf{B}_{\boldsymbol{\theta}}]_{ij} \triangleq [E_p \{ \nabla_{\boldsymbol{\theta}} \ln f_X(\mathbf{x}_m; \boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}}^T \ln f_X(\mathbf{x}_m; \boldsymbol{\theta}) \}]_{ij} = E_p \left\{ \frac{\partial \ln f_X(\mathbf{x}_m; \boldsymbol{\theta})}{\partial \theta_i} \cdot \frac{\partial \ln f_X(\mathbf{x}_m; \boldsymbol{\theta})}{\partial \theta_j} \right\}. \quad (4.4)$$

**Lemma 4.1** Let  $\mathcal{F} = \{f_X | f_X(\mathbf{x}; \boldsymbol{\theta}) \text{ is a p.d.f. } \forall \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d\}$  be a family of parametric pdfs which is regular w.r.t. the pdf  $p_X(\mathbf{x})$ . Then:

- i. The function  $\zeta(\boldsymbol{\theta}) = \int \ln f_X(\mathbf{x}_m; \boldsymbol{\theta}) p_X(\mathbf{x}_m) d\mathbf{x}_m$  is finite and twice continuously differentiable on  $\Theta$ , and for every  $\boldsymbol{\theta} \in \Theta$ :

$$\frac{\partial \zeta(\boldsymbol{\theta})}{\partial \theta_i} = \int \frac{\partial \ln f_X(\mathbf{x}_m; \boldsymbol{\theta})}{\partial \theta_i} p_X(\mathbf{x}_m) d\mathbf{x}_m < \infty \quad i = 1, \dots, p, \quad (4.5)$$

$$\frac{\partial^2 \zeta(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} = [\mathbf{A}_{\boldsymbol{\theta}}]_{ij} < \infty \quad i, j = 1, \dots, p \quad (4.6)$$

Moreover,  $[\mathbf{B}_{\boldsymbol{\theta}}]_{ij} < \infty$  for  $i, j = 1, \dots, p$ , and  $\mathbf{A}_{\boldsymbol{\theta}_0}$  is negative definite, where  $\boldsymbol{\theta}_0$  is the pseudo-true parameters vector defined in Eq. (4.2).

- ii. If  $p_X(\mathbf{x}) = f_X(\mathbf{x}; \bar{\boldsymbol{\theta}})$  for some  $\bar{\boldsymbol{\theta}} \in \Theta$ , then  $\boldsymbol{\theta}_0 = \bar{\boldsymbol{\theta}}$  and

$$\mathbf{A}_{\boldsymbol{\theta}_0} + \mathbf{B}_{\boldsymbol{\theta}_0} = \mathbf{0}. \quad (4.7)$$

The proof of this lemma can be found in Ref. [5]. In particular, Eq. (4.7) represents the classical equivalence between the two expressions of the Fisher information matrix (FIM) under correct model specification.

#### 4.4.2 MS-UNBIASED ESTIMATORS AND THE MCRB

Upon setting the necessary regularity conditions, a covariance inequality in the presence of misspecified regular models can be defined. First, the concept of *misspecified unbiasedness*, in short MS-unbiasedness, has to be briefly introduced.

**Definition 4.2 MS-unbiasedness [5]** Let  $\mathcal{P}$  be a family of pdfs and assume that the (misspecified) parametric model  $\mathcal{F}$  is regular w.r.t.  $\mathcal{P}$ . Let  $\mathbf{g}(\cdot)$  be a continuously differentiable mapping from  $\Theta$  to  $\Phi \subset \mathbb{R}^s$ . Let  $\boldsymbol{\varphi}(\mathbf{x})$  be a statistic from the iid observations  $\mathbf{x} = \{\mathbf{x}_m\}_{m=1}^M$  that takes its values in  $\Phi$ . Then,  $\boldsymbol{\varphi}(\mathbf{x})$  is an MS-unbiased estimator of  $\mathbf{g}(\boldsymbol{\theta}_0)$ , derived under the misspecified model  $\mathcal{F}$ , iff:

$$E_p\{\boldsymbol{\varphi}(\mathbf{x})\} = \int \boldsymbol{\varphi}(\mathbf{x}) p_X(\mathbf{x}) d\mathbf{x} = \mathbf{g}(\boldsymbol{\theta}_0), \quad \forall p_X(\mathbf{x}) \in \mathcal{P}. \quad (4.8)$$

As in the classical estimation framework, the function  $\mathbf{g}(\cdot)$  is introduced in order to take into account all cases in which one can be interested in subsets or, more generally, in a given transformation of the (pseudo-true) parameter vector.

It is easy to show that the previous definition is consistent with the classical definition of unbiasedness. Without lack of generality, assume that  $\mathbf{g}$  is an identity mapping, i.e.,  $\mathbf{g}(\boldsymbol{\theta}) = \boldsymbol{\theta}$ . Then, the statistic  $\boldsymbol{\varphi}(\mathbf{x})$  is exactly an estimator of the pseudo-true parameter vector  $\boldsymbol{\theta}_0$ ; so, it is reasonable to use the standard notation  $\boldsymbol{\varphi}(\mathbf{x}) \triangleq \hat{\boldsymbol{\theta}}(\mathbf{x})$ . When the model  $\mathcal{F}$  is correctly specified, there exists a  $\bar{\boldsymbol{\theta}} \in \Theta$  such that  $p_X(\mathbf{x}) = f_X(\mathbf{x}; \bar{\boldsymbol{\theta}})$  for every  $\mathbf{x}$ . Then, from Lemma 4.1,  $\bar{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$  and finally Eq. (4.8) reduces to  $E_{f_{\bar{\boldsymbol{\theta}}}}\{\hat{\boldsymbol{\theta}}(\mathbf{x})\} = \int \hat{\boldsymbol{\theta}}(\mathbf{x}) f_X(\mathbf{x}; \bar{\boldsymbol{\theta}}) d\mathbf{x} = \bar{\boldsymbol{\theta}}$  that is exactly the standard definition of unbiasedness.

At this point, a lower bound in the presence of (regular) misspecified models can be introduced.

**Theorem 4.1 The misspecified Cramér-Rao bound, MCRB [5,20]** Let  $\mathcal{F}$  be a parametric model. Let  $\mathcal{P}(\mathcal{F})$  be the family of all pdfs w.r.t. which  $\mathcal{F}$  is regular. Suppose that  $\mathcal{P}(\mathcal{F})$  is not empty. Let  $\mathbf{g}(\cdot)$  be a continuously differentiable mapping from  $\Theta$  to  $\Phi \subset \mathbb{R}^s$ . Let  $\boldsymbol{\varphi}(\mathbf{x})$  be an MS-unbiased estimator derived under the misspecified model  $\mathcal{F}$  from the iid observed vectors  $\mathbf{x} = \{\mathbf{x}_m\}_{m=1}^M$ . Then, for every  $p_X(\mathbf{x})$  in  $\mathcal{P}(\mathcal{F})$ :

$$\mathbf{C}(\boldsymbol{\varphi}(\mathbf{x}), \mathbf{g}(\boldsymbol{\theta}_0)) \geq \frac{1}{M} \mathbf{G}_{\boldsymbol{\theta}_0} \mathbf{A}_{\boldsymbol{\theta}_0}^{-1} \mathbf{B}_{\boldsymbol{\theta}_0} \mathbf{A}_{\boldsymbol{\theta}_0}^{-1} \mathbf{G}_{\boldsymbol{\theta}_0}^T \triangleq \text{MCRB}(\boldsymbol{\theta}_0), \quad (4.9)$$

where

$$\mathbf{C}_p(\boldsymbol{\varphi}(\mathbf{x}), \mathbf{g}(\boldsymbol{\theta}_0)) \triangleq E_p\left\{ (\boldsymbol{\varphi}(\mathbf{x}) - \mathbf{g}(\boldsymbol{\theta}_0)) (\boldsymbol{\varphi}(\mathbf{x}) - \mathbf{g}(\boldsymbol{\theta}_0))^T \right\} \quad (4.10)$$

is the error covariance matrix of  $\boldsymbol{\varphi}(\mathbf{x})$ , the matrices  $\mathbf{A}_{\boldsymbol{\theta}_0}$  and  $\mathbf{B}_{\boldsymbol{\theta}_0}$  are defined in Eqs. (4.1) and (4.4), respectively, and  $\mathbf{G}_{\boldsymbol{\theta}_0} = \nabla_{\boldsymbol{\theta}_0}^T \mathbf{g}(\boldsymbol{\theta}_0)$  is the Jacobian matrix of  $\mathbf{g}$  evaluated at  $\boldsymbol{\theta}_0$ . Following Ref. [13], we refer to the right side of Eq. (4.9) as the MCRB.

The proof of this theorem can be found in Ref. [5]. It can be noted that the hypothesis that  $\mathcal{P}(\mathcal{F})$  is not empty is not so strong. In fact, it requires that there exists at least one pdf  $p_X(\mathbf{x}_m)$  for which, from Assumption 4.2, the point  $\boldsymbol{\theta}_0$  exists [5]. In the following, we provide many examples in which it is possible to evaluate  $\boldsymbol{\theta}_0$  and so the MCRB applies. Other relevant signal processing problems in which the pseudo-true vector  $\boldsymbol{\theta}_0$  can be evaluated are discussed in Refs. [13,15].

It is worth noting that the MCRB is consistent with the classical CRB. As for the unbiasedness, assuming for simplicity that  $\mathbf{g}(\boldsymbol{\theta}) = \boldsymbol{\theta}$ , then  $\boldsymbol{\varphi}(\mathbf{x}) \triangleq \hat{\boldsymbol{\theta}}(\mathbf{x})$ , when the model  $\mathcal{F}$  is correctly specified,  $p_X(\mathbf{x}) = f_X(\mathbf{x}; \bar{\boldsymbol{\theta}})$  for some  $\bar{\boldsymbol{\theta}} \in \Theta$ . Then, from Lemma 4.1, the matrices  $-\mathbf{A}_{\bar{\boldsymbol{\theta}}}$  and  $\mathbf{B}_{\bar{\boldsymbol{\theta}}}$  are equal and correspond to the classical FIM, and finally:

$$\begin{aligned}\mathbf{C}_{f_0}(\hat{\boldsymbol{\theta}}(\mathbf{x}), \bar{\boldsymbol{\theta}}) &\geq \frac{1}{M} \mathbf{A}_{\bar{\boldsymbol{\theta}}}^{-1} \mathbf{B}_{\bar{\boldsymbol{\theta}}} \mathbf{A}_{\bar{\boldsymbol{\theta}}}^{-1} = -\frac{1}{M} \mathbf{A}_{\bar{\boldsymbol{\theta}}}^{-1} = \frac{1}{M} \mathbf{B}_{\bar{\boldsymbol{\theta}}} = -\frac{1}{M} \left( E_{f_0} \left\{ \nabla_{\bar{\boldsymbol{\theta}}} \nabla_{\bar{\boldsymbol{\theta}}}^T \ln f_X(\mathbf{x}_m; \bar{\boldsymbol{\theta}}) \right\} \right)^{-1} \\ &= \frac{1}{M} \left( E_{f_0} \left\{ \nabla_{\bar{\boldsymbol{\theta}}} \ln f_X(\mathbf{x}_m; \bar{\boldsymbol{\theta}}) \nabla_{\bar{\boldsymbol{\theta}}}^T \ln f_X(\mathbf{x}_m; \bar{\boldsymbol{\theta}}) \right\} \right)^{-1},\end{aligned}\quad (4.11)$$

which represents the classical Cramér-Rao inequality for any unbiased estimator.

*Remark 4.1.* The statement and the proof of Theorem 4.1, given in Ref. [5], consider only the case of *real* parameter space, i.e.,  $\Theta \subset \mathbb{R}^d$ . However, as shown in Refs. [13, 76], the derivation can be easily extended to the complex case, i.e., when  $\Theta \subset \mathbb{C}^d$ . This is because all the pdfs are *real functions of complex variables* ( $\mathbf{x}$  and  $\boldsymbol{\theta}$ ), so we do not need sophisticated holomorphic calculus to generalize the derivatives w.r.t. a complex parameter vector  $\boldsymbol{\theta}$ . Insightful procedures, useful to generalize the Cramér-Rao inequality in the complex case, are discussed in Refs. [26–28].

*Remark 4.2.* In order to evaluate the MCRB of Eq. (4.9), the knowledge of the true pdf  $p_X(\mathbf{x})$  is required. However, this should not be seen as a limitation of its applicability. Think for example of the common situation in which one knows that the true data distribution is given by an involved function that does not admit an easy analytical tractability, e.g., the rendering of the ML estimator is difficult or impossible to derive. In these cases, one typically assumes a simpler model, such as a Gaussian model, introducing a mismatch. The evaluation of the MCRB would show the potential performance loss due to the mismatch between the assumed and the true model. An example of this procedure is discussed in this chapter, in the context of the scatter matrix estimation problem for radar detection applications. Another useful application of the MCRB is the prediction of possible weaknesses (i.e., breakdown of the estimation performance) of the system under uncommon conditions. In particular, given an assumed model for the data, one can be interested in evaluating the performance loss in the presence of a certain number of “true” possible data distributions that the system can undergo.

We note, in passing, that in all the situations in which the true pdf is known but it is not possible to evaluate, in closed form, the expectation operator involved in the definition of the matrices  $\mathbf{A}_{\boldsymbol{\theta}_0}$  in Eq. (4.1) and  $\mathbf{B}_{\boldsymbol{\theta}_0}$  in Eq. (4.4), the MCRB can be approximated by means of Monte Carlo simulations.

*Remark 4.3.* Before introducing the MML estimator, we briefly comment on the difference between the results obtained in Ref. [13] and the general derivation of the MCRB provided in Ref. [5]. Even if the final mismatched covariance inequality assumes exactly the same expression (at least when  $\mathbf{g}$  is an identity mapping and then

$\Phi(\mathbf{x}) \triangleq \hat{\boldsymbol{\theta}}(\mathbf{x})$ ), the proof in Ref. [5] is general, while the one provided in Ref. [13] relies on first-order Taylor expansion of the estimation error (see Eq. 41 in Ref. [13]). This derivation leads to define a restricted class of estimators for which the MCRB of Eq. (4.9) applies which is defined by the following two properties:

1. The expected value w.r.t. the true distribution is the same for all the estimators in the class and is equal to  $E_p\{\hat{\boldsymbol{\theta}}(\mathbf{x})\} = \boldsymbol{\mu}$ ,
2. The correlation matrix  $\boldsymbol{\Xi}_{\boldsymbol{\theta}}$  between the estimation error and the score function  $\boldsymbol{\eta}_{\boldsymbol{\theta}}(\mathbf{x})$ , i.e.,

$$\boldsymbol{\Xi}_{\boldsymbol{\theta}} = E_p\left\{ (\hat{\boldsymbol{\theta}}(\mathbf{x}) - \boldsymbol{\mu}) \boldsymbol{\eta}_{\boldsymbol{\theta}}(\mathbf{x})^T \right\} \quad (4.12)$$

must be equal to some matrix function  $\mathbf{M}(\boldsymbol{\theta})$ , such that  $\boldsymbol{\Xi}_{\boldsymbol{\theta}} = \pm \mathbf{M}(\boldsymbol{\theta})$  for all the estimators in the class. The score function used in Ref. [13] and in Ref. [19] is  $\boldsymbol{\eta}_{\boldsymbol{\theta}}(\mathbf{x}) = \nabla_{\boldsymbol{\theta}} \ln f_X(\mathbf{x}; \boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}} D(p \| f_{\boldsymbol{\theta}})$ . For this function, it turns out that the correlation matrix  $\boldsymbol{\Xi}_{\boldsymbol{\theta}}$  must be equal to  $\pm \mathbf{A}_{\boldsymbol{\theta}_0}^{-1} \mathbf{B}_{\boldsymbol{\theta}_0}$ . This particular choice of the score function has been motivated by its “tightness” properties discussed in Ref. [29].

As shown in Ref. [13], there is at least an estimator that *asymptotically* satisfies constraints (1) and (2). This estimator is exactly the MML estimator described in the next section. However, in general, it would be very difficult to characterize explicitly a class of estimators that satisfy these two constraints. The advantage of the proof in Ref. [5] is that it shows that inequality (4.9) holds for all the MS-unbiased estimators and not only for those satisfying (1) and (2).

#### 4.4.3 THE MISMATCHED MAXIMUM LIKELIHOOD (MML) ESTIMATOR

In the previous subsection, the MCRB has been introduced. In particular, we showed that it represents the counterpart of the classical CRB in the presence of model misspecification. At this point, a question that naturally arises is to whether there exists a mismatched estimator whose error covariance matrix is equal (at least asymptotically) to the MCRB. As we will see, the answer is “yes” and this “misspecified-efficient” estimator is a generalization of the classical maximum likelihood (ML) estimator, called the MML estimator or also the QML estimator [2]. In the rest of this section, we assume that  $\mathbf{g}$  is an identical mapping so that  $\mathbf{G}_{\boldsymbol{\theta}} \triangleq \nabla_{\boldsymbol{\theta}}^T \mathbf{g}(\boldsymbol{\theta}) = \mathbf{I}, \forall \boldsymbol{\theta} \in \Theta$  where  $\mathbf{I}$  is the identity matrix.

The MML estimator has been introduced in Ref. [1,2] as:

$$\hat{\boldsymbol{\theta}}_{MML}(\mathbf{x}) = \arg \max_{\boldsymbol{\theta} \in \Theta} \ln f_X(\mathbf{x}; \boldsymbol{\theta}) = \arg \max_{\boldsymbol{\theta} \in \Theta} \sum_{m=1}^M \ln f_X(\mathbf{x}_m; \boldsymbol{\theta}), \quad (4.13)$$

where  $\mathbf{x}_m \sim p_X(\mathbf{x}_m)$ . It can be shown (see [1] and [2]) that the MML estimator converges *almost surely* (*a.s.*) to the  $\boldsymbol{\theta}_0$  introduced in Eq. (4.2), i.e., the vector that minimizes the KL divergence between  $p_X(\mathbf{x}_m)$  and  $f_X(\mathbf{x}_m; \boldsymbol{\theta})$ :

$$\hat{\boldsymbol{\theta}}_{MML}(\mathbf{x}) \xrightarrow[M \rightarrow \infty]{a.s.} \boldsymbol{\theta}_0, \quad (4.14)$$

Under similar regularity conditions to those given in [Section 4.4.1 \(Assumptions 4.1–4.3\)](#), the asymptotic normality of the MML estimator is proved in Refs. [1,2]. This result is summarized by the following theorem (see Refs. [1,2] for the proof):

**Theorem 4.2** [1,2] *Under suitable regularity conditions, it can be proved that*

$$\sqrt{M}(\hat{\boldsymbol{\theta}}_{MML}(\mathbf{x}) - \boldsymbol{\theta}_0) \xrightarrow[M \rightarrow \infty]{d} \mathcal{N}\left(\mathbf{0}, \mathbf{A}_{\boldsymbol{\theta}_0}^{-1} \mathbf{B}_{\boldsymbol{\theta}_0} \mathbf{A}_{\boldsymbol{\theta}_0}^{-1}\right), \quad (4.15)$$

where  $\xrightarrow[M \rightarrow \infty]{d}$  indicates the convergence in distribution and the matrices  $\mathbf{A}_{\boldsymbol{\theta}_0}$  and  $\mathbf{B}_{\boldsymbol{\theta}_0}$  have been defined in Eqs. (4.1) and (4.4), respectively. The asymptotic covariance matrix  $\mathbf{A}_{\boldsymbol{\theta}_0}^{-1} \mathbf{B}_{\boldsymbol{\theta}_0} \mathbf{A}_{\boldsymbol{\theta}_0}^{-1}$  is generally called Huber's "sandwich" covariance.

[Theorems 4.1](#) and [4.2](#) highlight an important fact: the MML estimator is asymptotically MS-unbiased and its error covariance matrix asymptotically equates the MCRB. The similarity with the classical (matched) estimation framework is now clear: the MML estimator is the counterpart of the ML estimator in the presence of misspecified models, as the MCRB is the counterpart of the classical (matched) CRB. However, it must be noted that while in the classical matched case, the convergence and the unbiasedness of the ML estimator is defined w.r.t. the true parameter vector, in the mismatched case the convergence and the MS-unbiasedness of the MML estimator is always defined w.r.t. the pseudo-true parameter vector  $\boldsymbol{\theta}_0$  of Eq. (4.2). The next subsection provides some insights about this important fact.

#### 4.4.4 A PARTICULAR CASE: THE MCRB AS A BOUND ON THE MEAN SQUARE ERROR (MSE)

In this section, we focus on a particular mismatched case: the unknown parameter space  $T$  of the true model is the same of the parameter space  $\Theta$  of the assumed model, i.e.,  $T \equiv \Theta \subset \mathbb{R}^d$  [20]. As before, we assume that  $\mathbf{g}$  is an identical mapping so that  $\boldsymbol{\varphi}(\mathbf{x}) \triangleq \hat{\boldsymbol{\theta}}(\mathbf{x})$ . More formally, assume that the true parametric pdf of the observations  $p_X(\mathbf{x})$  and the assumed pdf  $f_X(\mathbf{x}; \boldsymbol{\theta})$  belong to two (generally different) families of pdf's,  $\mathcal{P}$  and  $\mathcal{F}$  that can be parameterized by using the same parameter space  $\Theta$ :

$$\mathcal{P} = \{p_{\boldsymbol{\theta}} | p_X(\mathbf{x}; \boldsymbol{\theta}) \text{ is a p.d.f. } \forall \boldsymbol{\theta} \in \Theta\}, \quad \mathcal{F} = \{f_{\boldsymbol{\theta}} | f_X(\mathbf{x}; \boldsymbol{\theta}) \text{ is a p.d.f. } \forall \boldsymbol{\theta} \in \Theta\}. \quad (4.16)$$

An application in which this particular case arises will be discussed in [Section 4.6.1](#) of this chapter. Even if this is only a particular case of the theory developed in the previous sections, this type of mismatch allows us to deeply understand the nature of the MCRB and of the MML estimator. In particular, if condition (4.16) is satisfied, we can directly compare the MCRB and the MML estimator with their classical (matched) counterparts, i.e., the CRB and the ML estimator. This can be done since the pseudo-true parameter vector  $\boldsymbol{\theta}_0$  belongs to the same parameter space of the true parameter vector  $\bar{\boldsymbol{\theta}}$ , and as such the difference vector  $\mathbf{r} \triangleq \bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_0$  is well defined. In essence, the vector  $\mathbf{r}$  is in general different from a zero vector.

In particular, vector  $\mathbf{r}$  indicates the distance between the convergence point  $\bar{\boldsymbol{\theta}}$  of the classical ML estimator when the true pdf  $p_X(\mathbf{x}; \boldsymbol{\theta})$  is perfectly known and the convergence point  $\boldsymbol{\theta}_0$  of the MML estimator when the mismatched pdf  $f_X(\mathbf{x}; \boldsymbol{\theta})$  which satisfies the condition in (4.16) is adopted. Moreover, using  $\mathbf{r}$ , a bound on the *mean square error* (MSE) of the estimate of  $\bar{\boldsymbol{\theta}}$  in the presence of mismatched models can be easily established. To proceed, the error covariance matrix of Eq. (4.10) can be rewritten as:

$$\begin{aligned} \text{MSE}_p(\hat{\boldsymbol{\theta}}(\mathbf{x}), \bar{\boldsymbol{\theta}}) &\triangleq E_p \left\{ (\hat{\boldsymbol{\theta}}(\mathbf{x}) - \boldsymbol{\theta}_0 + \boldsymbol{\theta}_0 - \bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}(\mathbf{x}) - \boldsymbol{\theta}_0 + \boldsymbol{\theta}_0 - \bar{\boldsymbol{\theta}})^T \right\} \\ &= \mathbf{C}_p(\hat{\boldsymbol{\theta}}(\mathbf{x}), \boldsymbol{\theta}_0) - 2E_p \left\{ \hat{\boldsymbol{\theta}}(\mathbf{x}) - \boldsymbol{\theta}_0 \right\} (\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T + (\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)(\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^T \\ &= \mathbf{C}_p(\hat{\boldsymbol{\theta}}(\mathbf{x}), \boldsymbol{\theta}_0) + \mathbf{r}\mathbf{r}^T, \end{aligned} \quad (4.17)$$

where  $E_p \left\{ \hat{\boldsymbol{\theta}}(\mathbf{x}) - \boldsymbol{\theta}_0 \right\} = 0$  due to the MS-unbiasedness assumption. A similar expansion of the MSE can be found in Ref. [13]. Finally, by substituting the covariance inequality in (4.9) in Eq. (4.17), we can obtain a misspecified bound on the MSE of  $\bar{\boldsymbol{\theta}}$  as:

$$\text{MSE}_p(\hat{\boldsymbol{\theta}}(\mathbf{x}), \bar{\boldsymbol{\theta}}) \geq \frac{1}{M} \mathbf{A}_{\boldsymbol{\theta}_0}^{-1} \mathbf{B}_{\boldsymbol{\theta}_0} \mathbf{A}_{\boldsymbol{\theta}_0}^{-1} + \mathbf{r}\mathbf{r}^T. \quad (4.18)$$

Moreover, if the condition in Eq. (4.16) is satisfied, the concept of consistency can be extended also to MS-unbiased mismatched estimators. In particular, we define as *consistent* an MS-unbiased mismatched estimator if, as the number of data vectors  $M$  goes to infinity, it converges a.s. to the *true* parameter vector  $\bar{\boldsymbol{\theta}}$ , i.e.,  $\hat{\boldsymbol{\theta}}(\mathbf{x}) \xrightarrow[M \rightarrow \infty]{a.s.} \boldsymbol{\theta}_0 = \bar{\boldsymbol{\theta}}$ .

The mismatched MSE inequality in (4.18) and the concept of consistency for MS-unbiased mismatched estimators can be useful to compare in a very intuitive and self-explicative manner the nature of the MCRB and of the MML estimator, discussed in both Sections 4.5 and 4.6.

#### 4.4.5 THE CONSTRAINED MCRB: CMCRB

In some applications, one has to deal with additional constraints on the unknown parameter vector that need to be satisfied by an estimation algorithm. To this end, a constrained version of the classical (matched) CRB has been proposed in Ref. [30]. Successive generalizations can be found in Ref. [31–33]. The aim of this section is to show that a generalization of the results obtained for the constrained CRB (CCRB) to the mismatched framework is also possible and a constrained version of the MCRB, i.e., the constrained MCRB (CMCRB) under a set of equality constraints on the parameter vector, can be derived. This was accomplished in a recent work [34] by the generalization of the procedure described in Ref. [33]. A local bijective transformation of the unknown parameter vector in a lower dimensional parameter space is exploited to build the constraints in the CRB derivation.

Let  $\hat{\boldsymbol{\theta}}(\mathbf{x})$  be a  $d$ -dimensional *MS*-unbiased (in a proper sense discussed later) estimator derived under the misspecified model  $\mathcal{F}$ . Suppose that  $\hat{\boldsymbol{\theta}}(\mathbf{x})$  is required to satisfy  $k$  (with  $k < d$ ) continuously differentiable constraints [32]:

$$\mathbf{f}(\hat{\boldsymbol{\theta}}(\mathbf{x})) = \mathbf{0}. \quad (4.19)$$

The  $k \times d$  Jacobian matrix of the constraints,  $\mathbf{F}_{\boldsymbol{\theta}} = \nabla_{\boldsymbol{\theta}}^T \mathbf{f}(\boldsymbol{\theta})$  is assumed to have full rank for any  $\boldsymbol{\theta} \in \Theta$  satisfying Eq. (4.19). Then there exists a matrix  $\mathbf{U} \in \mathbb{R}^{d \times (d-k)}$  whose columns form an orthonormal basis for the null space of  $\mathbf{F}_{\boldsymbol{\theta}}$ , that is:

$$\mathbf{F}_{\boldsymbol{\theta}} \mathbf{U} = \mathbf{0}, \quad \mathbf{U}^T \mathbf{U} = \mathbf{I}. \quad (4.20)$$

The matrix  $\mathbf{U}$  can be obtained numerically by calculating the  $d-k$  orthonormal eigenvectors associated with the zero eigenvalue of  $\mathbf{F}_{\boldsymbol{\theta}}$ .

Using this framework, Stoica and Ng [32] derived a constrained version of the classical CRB. A different, yet equivalent, approach was adopted in Ref. [33], where the authors exploited the Implicit Function Theorem (see, e.g., [35, Th. 5-2]) to obtain the same CCRB of Ref. [32], but from a different standpoint. The starting point of the proof in Ref. [33] is that the constraint function  $\mathbf{f}$  restricts  $\boldsymbol{\theta}$  to a manifold  $\overline{\Theta} = \{\boldsymbol{\theta} | \mathbf{f}(\boldsymbol{\theta}) = \mathbf{0}\}$  of the original vector space  $\mathbb{R}^d$  with dimension  $d-k$  (since  $\text{rank}(\mathbf{F}_{\boldsymbol{\theta}}) = k$  by assumption). Therefore from the Implicit Function Theorem, there exist two open sets  $O$  and  $P$  of  $\overline{\Theta}$  and  $\mathbb{R}^{d-k}$ , respectively, and a local continuously differentiable bijection  $\mathbf{h}: P \ni \xi \rightarrow O \ni \boldsymbol{\theta}$  such that

$$\boldsymbol{\theta} = \mathbf{h}(\xi). \quad (4.21)$$

We denote as  $\mathbf{H}_{\xi} = \nabla_{\xi}^T \mathbf{h}(\xi)$  the  $d \times (d-k)$  full rank Jacobian matrix of the transformation  $\mathbf{h}$ . The idea behind the proof in Ref. [34] is the same as the one in Ref. [33]: given the MCRB, derived for an *intrinsic* parameter vector  $\xi_0$  belonging to the reduced parameter space  $\mathbb{R}^{d-k}$ , we can “project back” such bound on the manifold  $\overline{\Theta}$  by exploiting the bijective transformation  $\mathbf{h}$ . As we will see, the final expression of the CMCRB does not involve either the *intrinsic* parameter vector  $\xi_0$  or  $\mathbf{h}$ . In particular, the explicit knowledge of  $\xi_0$  and  $\mathbf{h}$  is not required, only their existence (guaranteed by the Implicit Function Theorem) is necessary. This is an important fact, since in general, to obtain an explicit expression for  $\mathbf{h}$  is not an easy task.

It is worth noticing that the constraints in Eq. (4.19) apply to  $\boldsymbol{\theta} \in \Theta$ , i.e., the parameter vector that parameterizes the assumed (and possibly misspecified) pdf  $f_X(\mathbf{x}; \boldsymbol{\theta})$ . They are not supposed to apply to  $\boldsymbol{\tau} \in T$ , i.e., the true and inaccessible parameter vector that in general may have, as discussed before, a completely different structure.

#### 4.4.5.1 The MCRB for the intrinsic parameter vector

As sketched in the previous subsection, the first step to derive the CMCRB is to analyze the conditions that guarantee the existence of the intrinsic pseudo-true parameter vector  $\xi_0$ . Then an explicit expression for the intrinsic MCRB for  $\xi_0$  is obtained by building upon the results discussed in Section 4.4.2.

### Existence of $\xi_0$

The assumptions needed to establish the existence of a pseudo-true intrinsic parameter vector  $\xi_0$  are formally the same as the [Assumptions 4.1–4.3](#) given in [Section 4.4.1](#). In order to improve the clarity of the presentation, we recall them here by specializing them to the constrained problem at hand. In particular, we assume that:

- i. The function  $\zeta(\boldsymbol{\theta}) = E_p \{ \ln f_X(\mathbf{x}_m; \boldsymbol{\theta}) | \boldsymbol{\theta} \in \bar{\Theta} \}$  has a unique maximum on  $\bar{\Theta}$  at an interior point  $\boldsymbol{\theta}_0$ , and hence  $\bar{\zeta}(\boldsymbol{\xi}) = E_p \{ \ln f_X(\mathbf{x}; \mathbf{h}(\boldsymbol{\xi})) | \boldsymbol{\xi} \in \mathbb{R}^{d-k} \}$  has a unique maximum on  $\mathbb{R}^{d-k}$  at a point  $\boldsymbol{\xi}_0$ ,
- ii. The matrix  $\mathbf{A}_{\xi_0}$  defined as:

$$\mathbf{A}_{\xi_0} = E_p \left\{ \nabla_{\boldsymbol{\xi}_0} \nabla_{\boldsymbol{\xi}_0}^T \ln f_X(\mathbf{x}; \mathbf{h}(\boldsymbol{\xi}_0)) \right\} \quad (4.22)$$

is nonsingular. This can be recognized also as the identifiability condition (see [22–25]) for  $\boldsymbol{\xi}_0$ . As before, the interior point  $\bar{\Theta} \ni \boldsymbol{\theta}_0 = \mathbf{h}(\boldsymbol{\xi}_0)$  can be equivalently seen as the point that minimizes the Kullback-Leibler divergence between the true distribution  $p_X(\mathbf{x})$  and the assumed distribution  $f_X(\mathbf{x}; \boldsymbol{\theta})$  with  $\boldsymbol{\theta} \in \bar{\Theta}$ , i.e.,  $\boldsymbol{\theta}_0 = \arg \min_{\boldsymbol{\theta} \in \bar{\Theta}} \{D(p_X || f_{\boldsymbol{\theta}})\}$ . This minimization problem can be rewritten as function of the intrinsic *pseudo-true parameter vector*  $\boldsymbol{\xi}_0$  as:

$$\boldsymbol{\xi}_0 = \arg \min_{\boldsymbol{\xi} \in \mathbb{R}^{d-k}} \{D(p_X || f_{\mathbf{h}(\boldsymbol{\xi})})\} = \arg \min_{\boldsymbol{\xi} \in \mathbb{R}^{d-k}} \{-E_p \{ \ln f_X(\mathbf{x}; \mathbf{h}(\boldsymbol{\xi})) \}\}. \quad (4.23)$$

### MS-unbiasedness and MCRB in $\boldsymbol{\xi}_0$

Under the two assumptions (i) and (ii) stated before, the definition of *MS-unbiasedness* and the expression of the MCRB given in Eqs. (4.8) and (4.9), respectively, can be easily exploited to derive the MCRB in  $\boldsymbol{\xi}_0$ . In particular, let  $\hat{\boldsymbol{\theta}}(\mathbf{x})$  be a constrained estimator derived under the misspecified model  $\mathcal{F}$  from the iid observations  $\mathbf{x} = \{\mathbf{x}_m\}_{m=1}^M$  that satisfies the set of equality constraints  $\mathbf{f}(\hat{\boldsymbol{\theta}}(\mathbf{x})) = \mathbf{0}$ . Let  $\mathbf{h}(\cdot)$  be the continuously differentiable bijective transformation in Eq. (4.21). Then,  $\hat{\boldsymbol{\theta}}(\mathbf{x})$  is an *MS-unbiased estimator* of  $\mathbf{h}(\boldsymbol{\xi}_0)$  iff:

$$E_p \left\{ \hat{\boldsymbol{\theta}}(\mathbf{x}) \right\} = \int \hat{\boldsymbol{\theta}}(\mathbf{x}) p_X(\mathbf{x}) d\mathbf{x} = \boldsymbol{\theta}_0 = \mathbf{h}(\boldsymbol{\xi}_0), \quad \forall p_X(\mathbf{x}) \in \mathcal{P}, \quad \boldsymbol{\theta}_0 \in \bar{\Theta}. \quad (4.24)$$

Moreover, the following inequality holds true:

$$\mathbf{C}_p \left( \hat{\boldsymbol{\theta}}(\mathbf{x}), \boldsymbol{\theta}_0 \right) = \mathbf{C}_p \left( \hat{\boldsymbol{\theta}}(\mathbf{x}), \mathbf{h}(\boldsymbol{\xi}_0) \right) \geq \frac{1}{M} \mathbf{H}_{\boldsymbol{\xi}_0} \mathbf{A}_{\boldsymbol{\xi}_0}^{-1} \mathbf{B}_{\boldsymbol{\xi}_0} \mathbf{A}_{\boldsymbol{\xi}_0}^{-1} \mathbf{H}_{\boldsymbol{\xi}_0}^T \triangleq \text{MCRB}(\boldsymbol{\xi}_0) \quad (4.25)$$

where  $\mathbf{C}_p \left( \hat{\boldsymbol{\theta}}(\mathbf{x}), \mathbf{h}(\boldsymbol{\xi}_0) \right) \triangleq E_p \left\{ \left( \hat{\boldsymbol{\theta}}(\mathbf{x}) - \mathbf{h}(\boldsymbol{\xi}_0) \right) \left( \hat{\boldsymbol{\theta}}(\mathbf{x}) - \mathbf{h}(\boldsymbol{\xi}_0) \right)^T \right\}$  is the error covariance matrix of  $\hat{\boldsymbol{\theta}}(\mathbf{x})$ , the matrix  $\mathbf{A}_{\boldsymbol{\xi}_0}$  is given in Eq. (4.22),  $\mathbf{B}_{\boldsymbol{\xi}_0}$  is defined as:

$$\mathbf{B}_{\boldsymbol{\xi}_0} \triangleq E_p \left\{ \nabla_{\boldsymbol{\xi}_0} \ln f_X(\mathbf{x}; \mathbf{h}(\boldsymbol{\xi}_0)) \nabla_{\boldsymbol{\xi}_0}^T \ln f_X(\mathbf{x}; \mathbf{h}(\boldsymbol{\xi}_0)) \right\}, \quad (4.26)$$

and  $\mathbf{H}_{\xi_0} = \nabla_{\xi_0}^T \mathbf{h}(\xi_0)$  is the Jacobian matrix of  $\mathbf{h}(\cdot)$  evaluated at  $\xi_0$ . The proof of the inequality in Eq. (4.25) follows directly by the proof of [Theorem 4.1](#) by substituting the invertible transformation  $\mathbf{g}(\cdot)$  with  $\mathbf{h}(\cdot)$ .

#### 4.4.5.2 The constrained MCRB (CMCRB)

Finally, from all previous results, an explicit expression for the CMCRB on  $\boldsymbol{\theta}_0$  is provided in the following theorem:

**Theorem 4.3 [34]** *The CMCRB is given by:*

$$\text{CMCRB}(\boldsymbol{\theta}_0) = \frac{1}{M} \mathbf{U} (\mathbf{U}^T \mathbf{A}_{\boldsymbol{\theta}_0} \mathbf{U})^{-1} \mathbf{U}^T \mathbf{B}_{\boldsymbol{\theta}_0} \mathbf{U} (\mathbf{U}^T \mathbf{A}_{\boldsymbol{\theta}_0} \mathbf{U})^{-1} \mathbf{U}^T, \quad \boldsymbol{\theta}_0 \in \overline{\Theta}, \quad (4.27)$$

where the (possibly singular) matrices  $\mathbf{A}_{\boldsymbol{\theta}_0}$  and  $\mathbf{B}_{\boldsymbol{\theta}_0}$  are defined, as in the unconstrained case, in Eqs. (4.1) and (4.4).

The proof of this theorem is given in Ref. [34] and an example of a possible application of the CMCRB to the scatter matrix estimation problem is provided in the following.

*Remark 4.4.* It is easy to verify that the CMCRB is consistent with the CCRB in Ref. [32]. When the model  $\mathcal{F}$  is correctly specified,  $p_X(\mathbf{x}) = f_X(\mathbf{x}; \bar{\boldsymbol{\theta}})$  for some  $\bar{\boldsymbol{\theta}} \in \overline{\Theta}$ . Then, the matrices  $-\mathbf{A}_{\bar{\boldsymbol{\theta}}}$  and  $\mathbf{B}_{\bar{\boldsymbol{\theta}}}$  are equal and correspond to the classical FIM, and finally:

$$\begin{aligned} \text{CMCRB}(\bar{\boldsymbol{\theta}}) &= \mathbf{U} (\mathbf{U}^T \mathbf{A}_{\bar{\boldsymbol{\theta}}} \mathbf{U})^{-1} \mathbf{U}^T \mathbf{B}_{\bar{\boldsymbol{\theta}}} \mathbf{U} (\mathbf{U}^T \mathbf{A}_{\bar{\boldsymbol{\theta}}} \mathbf{U})^{-1} \mathbf{U}^T \\ &= -\mathbf{U} (\mathbf{U}^T \mathbf{A}_{\bar{\boldsymbol{\theta}}} \mathbf{U})^{-1} \mathbf{U}^T = \mathbf{U} (\mathbf{U}^T \mathbf{B}_{\bar{\boldsymbol{\theta}}} \mathbf{U})^{-1} \mathbf{U}^T = \text{CCRB}(\bar{\boldsymbol{\theta}}), \end{aligned} \quad (4.28)$$

which represents the constrained CRB obtained in Ref. [32].

To conclude this section, it is worth noting that the derivation of an efficient constrained estimator able to achieve (at least asymptotically) the CMCRB is still an open problem. Moreover, the effects of inequality constraints need to be investigated as well.

---

## 4.5 TWO ILLUSTRATIVE EXAMPLES

In order to clarify the use of the MCRB and the MML estimator, two simple toy examples are described in the following. The problem is either to estimate the mean value ([Example 4.1](#)) [19] or the variance ([Example 4.2](#)) of Gaussian data [20]. Assume a set of  $M$  iid scalar observations  $\mathbf{x} = \{x_m\}_{m=1}^M$ , distributed according to a Gaussian pdf with mean value  $\mu_X$  and variance  $\sigma_X^2$ , i.e.,  $p_X(x_m) \equiv \mathcal{N}(\mu_X, \sigma_X^2)$ .

**Example 4.1** Estimation of the mean value:  $\bar{\theta} = \mu_X$

It is well known that, given the set of Gaussian observations  $\mathbf{x}$ , the ML estimator of the mean value is given by the *sample mean estimator*, i.e.,  $\hat{\theta}_{ML} = \frac{1}{M} \sum_{m=1}^M x_m$ . Suppose now that there is a mismatch in the assumed data variance. In other words, we assume for the data a Gaussian pdf  $f_X(x_m; \theta) = \mathcal{N}(\theta, \sigma^2)$  with a variance  $\sigma^2$  that is in general different from  $\sigma_X^2$ . It can be noted that in this simple example, the true unknown model  $p_X(x_m)$  and the assumed model  $f_X(x_m; \theta)$  admit the same parameterization, so this example falls in the particular case addressed in [Section 4.4.4](#). Following Eq. (4.13), a MML estimator for the mean value of the data is given by:

$$\hat{\theta}_{MML}(\mathbf{x}) = \arg \max_{\theta \in \Theta} \ln f_X(\mathbf{x}; \theta) = \arg \max_{\theta \in \Theta} \sum_{m=1}^M \ln f_X(x_m; \theta), \quad (4.29)$$

where

$$\ln f_X(x_m; \theta) = -\frac{1}{2} \ln 2\pi - \frac{1}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} (x_m - \theta)^2. \quad (4.30)$$

It is immediately clear that the MML estimator coincides with the ML estimator, i.e.,

$$\hat{\theta}_{MML}(\mathbf{x}) = \frac{1}{M} \sum_{m=1}^M x_m = \hat{\theta}_{ML}(\mathbf{x}). \quad (4.31)$$

Following the theory (see Eq. (4.14)), we know that the MML estimator converges a.s. to that point that minimizes the KL divergence between the true pdf  $p_X(x_m)$  and the assumed pdf  $f_X(x_m; \theta)$ . The KL divergence between  $p_X(x_m)$  and  $f_X(x_m; \theta)$  is given by Ref. [36]:

$$D(p_X \| f_\theta) = \frac{(\mu_X - \theta)^2}{2\sigma^2} + \frac{1}{2} \left( \frac{\sigma_X^2}{\sigma^2} - 1 - \ln \frac{\sigma_X^2}{\sigma^2} \right). \quad (4.32)$$

By taking the derivative with respect to  $\theta$  and by setting the result equal to 0, we obtain:

$$\frac{\partial D(p_X \| f_\theta)}{\partial \theta} \Big|_{\theta=\theta_0} = \frac{\mu_X - \theta}{\sigma^2} \Big|_{\theta=\theta_0} = 0, \quad (4.33)$$

whose solution is  $\theta_0 = \mu_X = \bar{\theta}$ . Eq. (4.33) shows that the MML converges a.s. to the true mean value and then, according to the definition provided in [Section 4.4.4](#), it is a *consistent estimator*. From the scalar version of Eq. (4.8), the mean value of the MML estimator w.r.t. the true joint pdf  $p_X(\mathbf{x})$  is:

$$E_p \{ \hat{\theta}_{MML}(\mathbf{x}) \} = \mu_X = \theta_0 = \bar{\theta}. \quad (4.34)$$

Hence, according to [Definition 4.1](#) given in [Section 4.4.2](#), the MML estimator is *MS-unbiased*. The MCRB can be evaluated as shown in Eq. (4.18) by evaluating the first and the second derivative of the  $\ln f_X(x_m; \theta)$  as:

$$\frac{\partial \ln f_X(x_m; \theta)}{\partial \theta} = -\frac{1}{\sigma^2}(x_m - \theta), \quad \frac{\partial^2 \ln f_X(x_m; \theta)}{\partial \theta^2} = \frac{1}{\sigma^2}. \quad (4.35)$$

In this case, matrices  $\mathbf{A}_{\theta_0}$  of Eq. (4.1) and  $\mathbf{B}_{\theta_0}$  of Eq. (4.4) are scalars and are easily derived to be:

$$A_{\theta_0} = E_p \left\{ \frac{1}{\sigma^2} \right\} = \frac{1}{\sigma^2}, \quad (4.36)$$

$$B_{\theta_0} = E_p \left\{ \frac{1}{\sigma^4} (x_k - \theta_0)^2 \right\} = \frac{1}{\sigma^4} E_p \left\{ (x_k - \mu_X)^2 \right\} = \frac{\sigma_X^2}{\sigma^4}. \quad (4.37)$$

Finally, from Eq. (4.18), we have that:

$$\text{MCRB}(\mu_X) = \sigma^2 \frac{\sigma_X^2}{M\sigma^4} \sigma^2 = \frac{\sigma_X^2}{M} = \text{CRLB}(\mu_X). \quad (4.38)$$

Note that, for a consistent estimator  $\mathbf{r} = \mathbf{0}$ . The fact that the MCRB and the CRB are equal is in accordance with the fact that the MML estimator is equal to the ML estimator and it does not depend on the misspecified variance  $\sigma^2$ .

**Example 4.2** Estimation of the variance:  $\bar{\theta} = \sigma_X^2$ .

In this example we consider the problem of estimating the variance of a Gaussian pdf in the presence of misspecified mean value, e.g., we erroneously assume that the mean value is zero when it is not. It is easy to show that, given the observation vector  $\mathbf{x}$ , the ML estimator of the variance is given by  $\hat{\theta}_{ML}(\mathbf{x}) = M^{-1} \sum_{m=1}^M (x_m - \mu_X)^2$ , where, as before,  $p_X(x_m) \equiv \mathcal{N}(\mu_X, \sigma_X^2)$ . Suppose now that the assumed Gaussian pdf is  $f_X(x_m; \theta) \equiv \mathcal{N}(\mu, \theta)$ , so we misspecify the mean value. As in Example 4.1, the true unknown model  $p_X(x_m)$  and the assumed model  $f_X(x_m; \theta)$  admit the same parameterization. Hence, also in this case, we can apply the results of Section 4.4.4. From Eq. (4.13), the MML estimator of the variance can be derived as:

$$\hat{\theta}_{MML}(\mathbf{x}) = \arg \max_{\theta \in \Theta} \ln f_X(\mathbf{x}; \theta) = \arg \max_{\theta \in \Theta} \sum_{m=1}^M \ln f_X(x_m; \theta), \quad (4.39)$$

where

$$\ln f_X(x_m; \theta) = -\frac{1}{2} \ln 2\pi - \frac{1}{2} \ln \theta - \frac{1}{2\theta} (x_m - \mu)^2. \quad (4.40)$$

It is easy to show that the MML estimator is given by:

$$\hat{\theta}_{MML}(\mathbf{x}) = \frac{1}{M} \sum_{m=1}^M (x_m - \mu)^2. \quad (4.41)$$

In this case, the KL divergence between  $p_X(x_m)$  and  $f_X(x_m; \theta)$  can be expressed as [36]:

$$D(p_X \| f_{\theta}) = \frac{(\mu_X - \mu)^2}{2\theta} + \frac{1}{2} \left( \frac{\sigma_X^2}{\theta} - 1 - \ln \frac{\sigma_X^2}{\theta} \right). \quad (4.42)$$

By taking the derivative w.r.t.  $\theta$  and by setting equal to zero the result, we get:

$$\frac{\partial D(p||f_\theta)}{\partial \theta} \Big|_{\theta=\theta_0} = \frac{-(\mu_X - \mu)^2 - \sigma_X^2}{2\theta^2} + \frac{1}{2\theta} \Big|_{\theta=\theta_0} = 0. \quad (4.43)$$

Hence,  $\theta_0 = \sigma_X^2 + (\mu_X - \mu)^2 \neq \bar{\theta}$ , i.e., the MML does not converge to the true variance, unless  $\mu = \mu_X$ , i.e., when there is no model mismatch. This means that the MML estimator of this example is not consistent. From the scalar version of Eq. (4.8), the mean value of the MML estimator with respect to the true distribution  $p_X(\mathbf{x}; \bar{\theta})$  is:

$$E_p\{\hat{\theta}_{MML}(\mathbf{x})\} = \sigma_X^2 + (\mu_X - \mu)^2 = \theta_0. \quad (4.44)$$

Hence, the MML estimator is *MS-unbiased* and the MCRB can be evaluated as shown in Eq. (4.18). By evaluating the first and the second derivative of the  $\ln f_X(x_m; \theta)$  and after some simple calculation, the matrices (that in this case are scalars)  $\mathbf{A}_{\theta_0}$  of Eq. (4.1) and  $\mathbf{B}_{\theta_0}$  Eq. (4.4) are obtained:

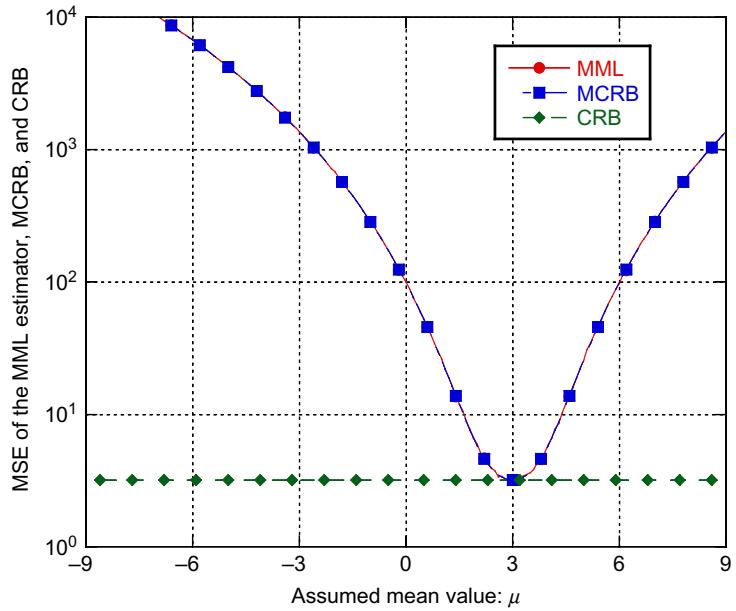
$$A_{\theta_0} = -\frac{1}{2\theta_0^2}, \quad B_{\theta_0} = \frac{3\sigma_X^4 + 6\sigma_X^2(\mu_X - \mu)^2 + (\mu_X - \mu)^4 - \theta_0^2}{4\theta_0^4}. \quad (4.45)$$

Finally, from Eq. (4.18), we have that:

$$\text{MCRB}(\bar{\theta}) = \text{MCRB}(\sigma_X^2) = \frac{2\sigma_X^4}{M} + \frac{4\sigma_X^2(\mu_X - \mu)^2}{M} + (\mu_X - \mu)^4. \quad (4.46)$$

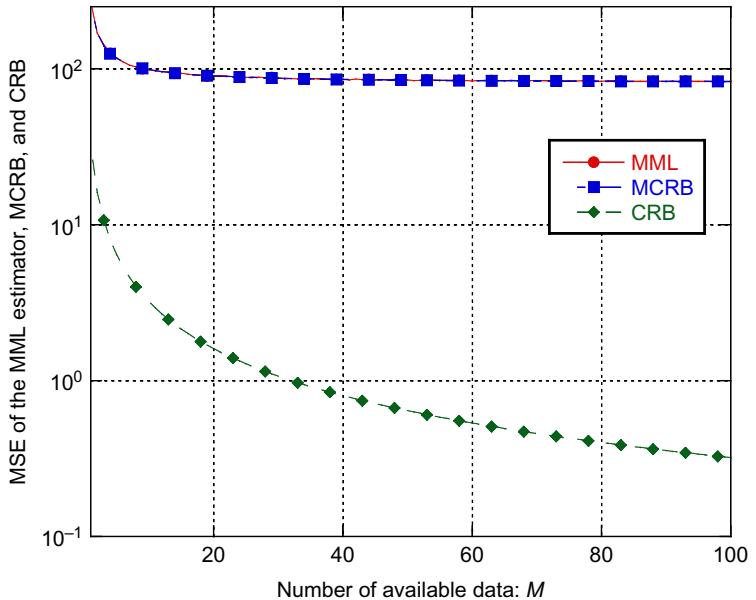
It is well known that the CRB for this estimation problem is given by  $\text{CRB}(\sigma_X^2) = 2\sigma_X^4/M$ . Hence, we obtain  $\text{MCRB}(\sigma_X^2) \geq \text{CRB}(\sigma_X^2)$ , i.e., for this estimation problem, the MCRB is always greater than or equal to the CRB. This is an intuitive result, since if we assume a wrong data model the best performance that we can achieve are worse than that in the case where we assume a correct data model. However, a general proof of an inequality between MCRB and CRB in the general case is still an open problem. When  $\mu = \mu_X$ , i.e., we correctly specify the mean value, then  $\theta_0 = \bar{\theta} = \sigma_X^2$  and  $\text{MCRB}(\sigma_X^2) = \text{CRB}(\sigma_X^2)$ .

In Figs. 4.1 and 4.2, we report the behavior of the square root of the MSE (RMSE) of the MML estimator in Eq. (4.41), the square root of the MCRB and of the CRLB, as function of the mismatched parameter (in this case the mean value  $\mu$ ) and as function of the number  $M$  of available data, respectively. In our simulation, the true mean value and the true variance are assumed to be  $\mu_X = 3$  and  $\sigma_X^2 = 4$ . As we can see from Fig. 4.1, the MCRB and the CRB are equal only when the true pdf and the assumed pdf are equal, i.e.,  $\mu = \mu_X = 3$  (the number of available data  $M$  is equal to 10). In Fig. 4.2, the RMSE, the MCRB, and the CRB are plotted as a function of the available number of data  $M$ . In this case, the assumed mean value is  $\mu = 0$ , hence there is some model mismatch. It is evident that the MCRB is a tight bound for the MML estimator, whereas the CRB is not.



**FIG. 4.1**

Comparison among the MSE of the MML estimator, the MCRB, and the CRB as function of the assumed mean value.



**FIG. 4.2**

Comparison among the MSE of the MML estimator, the MCRB, and the CRB as function of the available number of data  $M$ .

---

## 4.6 THE MCRB FOR THE ESTIMATION OF THE SCATTER MATRIX IN THE FAMILY OF CES DISTRIBUTIONS

In this section, we show the application of the MCRB theory to a realistic radar signal processing problem: the estimation of the disturbance covariance matrix from a set of acquired data vectors [20] (the so-called secondary data in the radar jargon). We recast this specific radar problem in the more general problem of estimating the  $N \times N$  scatter matrix of CES distributed data, given  $M$  iid realizations of the  $N$ -dimensional data vector  $\mathbf{x}_m$ , in the presence of data mismodeling. CES distributions constitute a wide family of distributions such as the complex Gaussian, Cauchy, generalized Gaussian, and compound Gaussian, which in turn includes the Gaussian distribution, the  $K$ -distribution, and the complex  $t$ -distribution [21]. The CES distributions, due to their flexibility, are widely applied in many areas, such as radar, sonar, and communications (see, e.g., [21,37–42]).

A complex  $N$ -dimensional random vector  $\mathbf{x}_m$  is CES distributed, in shorthand notation  $\mathbf{x}_m \sim CE_N(\boldsymbol{\gamma}, \boldsymbol{\Sigma}, h)$ , if its pdf is of the form:

$$p_X(\mathbf{x}_m) = c_{N,h} |\boldsymbol{\Sigma}|^{-1} h\left((\mathbf{x}_m - \boldsymbol{\gamma})^H \boldsymbol{\Sigma}^{-1} (\mathbf{x}_m - \boldsymbol{\gamma})\right), \quad (4.47)$$

where  $h$  is the density generator,  $c_{N,h}$  is a normalizing constant,  $\boldsymbol{\gamma} \triangleq E_p\{\mathbf{x}_m\}$ , and  $\boldsymbol{\Sigma}$  is the normalized (or shape) covariance matrix, also called *scatter matrix*. It is important to note that  $\boldsymbol{\gamma}$ ,  $\boldsymbol{\Sigma}$ , and  $h(\cdot)$  do not uniquely identify a CES distribution. In fact, given a CES distributed random vector  $\mathbf{x}_m \sim CE_N(\boldsymbol{\gamma}, \boldsymbol{\Sigma}, h)$ , for any  $\alpha > 0$ , we may define  $\tilde{\boldsymbol{\Sigma}} = \alpha \boldsymbol{\Sigma}$  and  $\tilde{h}(t) = h(t/\alpha)$  so that  $CE_N(\boldsymbol{\gamma}, \boldsymbol{\Sigma}, h) \equiv CE_N\left(\boldsymbol{\gamma}, \tilde{\boldsymbol{\Sigma}}, \tilde{h}\right)$  [21]. This ambiguity can be avoided by imposing a constraint on the scatter matrix  $\boldsymbol{\Sigma}$ , e.g.,  $\text{tr}(\boldsymbol{\Sigma}) = N$ , or by restricting the functional form of  $h(\cdot)$  in a suitable way. The difference between these two approaches is clarified in Sections 4.6.1 and 4.6.2.

Moreover, by imposing the constraint  $\text{tr}(\boldsymbol{\Sigma}) = N$ , if  $\mathbf{M} \triangleq E_p\left\{(\mathbf{x}_m - \boldsymbol{\gamma})(\mathbf{x}_m - \boldsymbol{\gamma})^H\right\}$  is the covariance matrix of the vector  $\mathbf{x}_m$ , then  $\boldsymbol{\Sigma} = (N/\text{tr}(\mathbf{M})) \cdot \mathbf{M}$ . It is important to observe that, for some CES distributions, the covariance matrix  $\mathbf{M}$  does not exist, but the scatter matrix  $\boldsymbol{\Sigma}$  is still well defined. Based upon the Stochastic Representation Theorem [21], any  $\mathbf{x}_m \sim CE_N(\boldsymbol{\gamma}, \boldsymbol{\Sigma}, h)$  with  $\text{rank}(\boldsymbol{\Sigma}) = k \leq N$  admits the stochastic representation  $\mathbf{x}_m = {}_d\boldsymbol{\gamma} + R\mathbf{T}\mathbf{u}$ , where the notation  $= {}_d$  means “has same distribution as.” The nonnegative random variable (r.v.)  $R \triangleq \sqrt{Q}$ , the so-called *modular variate*, is a real, nonnegative random variable,  $\mathbf{u}$  is a  $k$ -dimensional vector uniformly distributed on the unit hyper-sphere with  $k-1$  topological dimensions such that  $\mathbf{u}^H \mathbf{u} = 1$ ,  $R$  and  $\mathbf{u}$  are independent and  $\boldsymbol{\Sigma} = \mathbf{T}\mathbf{T}^H$  is a factorization of  $\boldsymbol{\Sigma}$ , where  $\mathbf{T}$  is a  $N \times k$  matrix and  $\text{rank}(\mathbf{T}) = k$ . In the following derivations, we assume that  $\boldsymbol{\Sigma}$  is full-rank, i.e.,  $\text{rank}(\mathbf{T}) = \text{rank}(\boldsymbol{\Sigma}) = N$ , and that it is real. For the CES distributions, the term  $\sigma_X^2 \triangleq E\{Q\}/N$  can be interpreted as the statistical power of

the random vector  $\mathbf{x}_m$ , i.e., the covariance matrix  $\mathbf{M}$  and the scatter matrix  $\Sigma$  are related by  $\mathbf{M} = \sigma_X^2 \Sigma$ . In general, the density generator itself depends on some additional parameters. For example, the complex  $t$ -distribution is completely characterized when its mean vector  $\gamma$ , scatter matrix  $\Sigma$ , shape parameter  $\lambda$ , and scale parameter  $\eta$  are perfectly known [21]. Since in many scenarios (e.g., radar and sonar) the disturbance vectors are zero mean, we set  $\gamma = \mathbf{0}$  in all the following analyses.

The application of the mismatched estimation framework to the problem of estimating the scatter matrix of a CES distributed random vector is described in two steps. First, we assume that all the characteristic parameters of the particular CES distributions at hands are *a priori* known, except for the elements of the scatter matrix  $\Sigma$ . Since, by definition, the scatter matrix is symmetric and of  $N \times N$  dimension, the parameters to be estimated are the  $l = N(N+1)/2$  elements of the lower (or upper) triangular submatrix of  $\Sigma$ . Then, the parameter vector that parameterizes a zero-mean CES distribution can be defined as  $\boldsymbol{\theta} = \text{vecs}(\Sigma)$ , where the  $\text{vecs}$ -operator is the “symmetric” counterpart of the standard  $\text{vec}$ -operator that maps a symmetric  $N \times N$  matrix  $\Sigma$  to a  $l$ -dimensional vector  $\boldsymbol{\theta}$  whose entries are the elements of the lower (or upper) triangular submatrix of  $\Sigma$  [43]. In this case, the ambiguity between the scatter matrix and the density generator  $h$  is removed by the assumption of the *a priori* knowledge of the extra parameters, i.e., we assume to know exactly the density generator. Consequently, the constraint on the trace of  $\Sigma$  is not required. It is worth noting that this is an unrealistic case, since in practical situations the *a priori* knowledge of the extra parameters which characterize the CES distributions is generally not available. However, this knowledge provides better understanding of how to apply the results on the MCRB and the MML estimator to this estimation problem. The more realistic case of unknown extra parameters is instead investigated in Section 4.6.2, where the problem of the joint estimation of the scatter matrix and of the extra parameters in the presence of misspecification is addressed.

#### 4.6.1 MISSPECIFIED ESTIMATION OF THE SCATTER MATRIX WITH PERFECTLY KNOWN EXTRA PARAMETERS

In the following, we assume that both the true distribution  $p_X(\mathbf{x}_m)$  and the assumed distribution  $f_X(\mathbf{x}_m; \boldsymbol{\theta})$  belong to the zero-mean CES distribution class:

$$p_X(\mathbf{x}_m) \triangleq p_X(\mathbf{x}_m; \bar{\Sigma}) = c_{N,h} |\bar{\Sigma}|^{-1} h\left(\mathbf{x}_m^H \bar{\Sigma}^{-1} \mathbf{x}_m\right), \quad (4.48)$$

$$f_X(\mathbf{x}_m; \boldsymbol{\theta}) \triangleq f_X(\mathbf{x}_m; \Sigma) = c_{N,g} |\Sigma|^{-1} g\left(\mathbf{x}_m^H \Sigma^{-1} \mathbf{x}_m\right), \quad (4.49)$$

where  $\bar{\boldsymbol{\theta}} = \text{vecs}(\bar{\Sigma})$ ,  $\boldsymbol{\theta} = \text{vecs}(\Sigma)$ ,  $h$  is the density generator of the true pdf, and  $g$  is the density generator of the assumed pdf. We propose three different case studies:

- *Case Study 1.* Assumed pdf: complex Normal; true pdf: *t*-student.
- *Case Study 2.* Assumed pdf: complex Normal, true pdf: Generalized Gaussian.
- *Case Study 3.* Assumed pdf: Generalized Gaussian; true pdf: *t*-student.

It can be noted that the true unknown pdf  $p_X(\mathbf{x}_m; \bar{\Sigma})$  and the assumed pdf  $f_X(\mathbf{x}_m; \Sigma)$  admit the same parameterization, so these examples fall in the particular case addressed in [Section 4.4.4](#).

#### 4.6.1.1 Case Study 1. Assumed pdf: complex Normal; true pdf: t-student.

We assume a complex Normal model for the data, i.e., each iid complex vector of the available dataset  $\mathbf{x} = \{\mathbf{x}_m\}_{m=1}^M$  is distributed according to a complex Normal multivariate pdf, which belongs to the CES family:

$$f_X(\mathbf{x}_m; \boldsymbol{\theta}) \triangleq f_X(\mathbf{x}_m; \Sigma) = \frac{1}{(\pi\sigma^2)^N |\Sigma|} \exp\left(-\frac{\mathbf{x}_m^H \Sigma^{-1} \mathbf{x}_m}{\sigma^2}\right). \quad (4.50)$$

The covariance matrix  $\mathbf{M} = E\{\mathbf{x}_m \mathbf{x}_m^H\} = \sigma^2 \Sigma$  in this case exists, provided that  $\sigma^2 < \infty$ . However, the true data are distributed according to another CES distribution, the complex *t*-distribution:

$$p_X(\mathbf{x}_m; \bar{\boldsymbol{\theta}}) \triangleq p_X(\mathbf{x}_m; \bar{\Sigma}) = \frac{1}{\pi^N |\bar{\Sigma}|} \frac{\Gamma(N+\lambda)}{\Gamma(\lambda)} \left(\frac{\lambda}{\eta}\right)^\lambda \left(\frac{\lambda}{\eta} + \mathbf{x}_m^H \bar{\Sigma}^{-1} \mathbf{x}_m\right)^{-(N+\lambda)}, \quad (4.51)$$

where  $\lambda$  is the shape parameter and  $\eta$  is the scale parameter characterizing the model [21,42]. The complex *t*-distribution has tails heavier than the Normal one for every  $\lambda \in (0, \infty)$ , while the limiting case  $\lambda \rightarrow \infty$  yields the complex Normal distribution.

The assumption of a complex Normal model is motivated by the fact that the MML estimator of the scatter matrix can be easily derived to be the well-known sample covariance matrix (SCM),  $\hat{\mathbf{M}}_{MML} = M^{-1} \sum_{m=1}^M \mathbf{x}_m \mathbf{x}_m^H$ , so [44]:

$$\hat{\mathbf{M}}_{MML} = \frac{\hat{\mathbf{M}}_{MML}}{\sigma^2} = \frac{1}{M\sigma^2} \sum_{m=1}^M \mathbf{x}_m \mathbf{x}_m^H, \quad (4.52)$$

where the power  $\sigma^2$  is assumed to be a priori known.

As first step, we evaluate the matrix that minimizes the KL divergence between  $p_X(\mathbf{x}_m)$ , considered here as a generic element of the CES family, and  $f_X(\mathbf{x}_m; \Sigma)$  (the complex Normal pdf). This matrix is the convergence point of the MML estimator in Eq. (4.52). The differential of the KL divergence with respect to  $\Sigma$  is given by [45]:

$$\begin{aligned} \partial D(p_X || f_\Sigma) &= -E_p\{\partial \ln f_X(\mathbf{x}_m; \Sigma)\} = -E_p\left\{\partial \ln |\Sigma|^{-1} + \partial \ln g(\mathbf{x}_m^H \Sigma^{-1} \mathbf{x}_m)\right\} \\ &= \text{tr}(\Sigma^{-1} \partial \Sigma) + \text{tr}\left(E_p\left\{\frac{d \ln g(Q_\Sigma)}{d Q_\Sigma} \Sigma^{-1} \mathbf{x}_m \mathbf{x}_m^H \Sigma^{-1} \partial \Sigma\right\}\right), \end{aligned} \quad (4.53)$$

where

$$Q_{\Sigma} \triangleq \mathbf{x}_m^H \Sigma^{-1} \mathbf{x}_m \quad (4.54)$$

The last equality in Eq. (4.53) follows directly from the same calculus given in Ref. [40,46]. Since the assumed distribution  $f_X(\mathbf{x}_m; \Sigma)$  is a complex Normal distribution, then  $g(Q_{\Sigma}) = \exp(-Q_{\Sigma}/\sigma^2)$  and  $d \ln g(Q_{\Sigma})/dQ_{\Sigma} = -1/\sigma^2$ . By substituting this result in Eq. (4.53), we get:

$$\partial D(p_X \| f_{\Sigma}) = \text{tr}(\Sigma^{-1} \partial \Sigma) - \frac{1}{\sigma^2} \text{tr}(E_p\{\Sigma^{-1} \mathbf{x}_m \mathbf{x}_m^H \Sigma^{-1} \partial \Sigma\}) = \text{tr}\left(\left[\Sigma^{-1} - \frac{\sigma_X^2}{\sigma^2} \Sigma^{-1} \bar{\Sigma} \Sigma^{-1}\right] \partial \Sigma\right), \quad (4.55)$$

where we used the property  $E_p\{\mathbf{x}_m \mathbf{x}_m^H\} = \sigma_X^2 \bar{\Sigma}$ . Then, following the standard rules of matrix calculus [45], the derivative of the KL divergence w.r.t.  $\Sigma$  is:

$$\frac{\partial}{\partial \Sigma} D(p_X \| f_{\Sigma}) = \Sigma^{-1} - \frac{\sigma_X^2}{\sigma^2} \Sigma^{-1} \bar{\Sigma} \Sigma^{-1}. \quad (4.56)$$

Finally, by setting the derivative in Eq. (4.56) equal to zero, we obtain matrix  $\Sigma_0$  that minimizes the KL divergence as:

$$\Sigma_0 = \frac{\sigma_X^2}{\sigma^2} \bar{\Sigma}. \quad (4.57)$$

Eq. (4.57) shows that the MML estimator converges *a.s.* to a scaled version of the true scatter matrix,  $\hat{\Sigma}_{MML}(\mathbf{x}) \xrightarrow[M \rightarrow \infty]{a.s.} \Sigma_0 = (\sigma_X^2 / \sigma^2) \bar{\Sigma}$ , so it is not consistent. It is consistent only when the two powers of the assumed and true pdfs are equal. The mean value of the MML estimator with respect to the true distribution is:

$$\boldsymbol{\mu} = E_p\{\hat{\Sigma}_{MML}(\mathbf{x})\} = \frac{\sigma_X^2}{\sigma^2} \bar{\Sigma} = \Sigma_0. \quad (4.58)$$

Hence, the MML estimator is *MS-unbiased*. Given the MS-unbiasedness of the proposed MML estimator, we can evaluate the MCRB. In Ref. [46], the MCRB on the estimation of the scatter matrix was evaluated for two CES distributions, the complex-*t* and the generalized Gaussian (GG), when the assumed misspecified distribution is a complex Normal pdf. In this case study, we assume that the true distribution is a complex-*t* distribution with pdf given in Eq. (4.51). The GG case will be discussed in the next case study.

Before providing the expression of the MCRB, some considerations on a reasonable choice of the true distribution parameters,  $\lambda$  and  $\eta$ , have to be made. The power  $\sigma_X^2 \triangleq E_p\{Q_{\Sigma}\}/N$  is function of these two parameters. By applying the Stochastic Representation Theorem, we have that  $Q_{\Sigma}$  has an *F*-distribution [21] such that

$$p_{Q_{\Sigma}}(q) = \frac{1}{B(N, \lambda)} q^{N-1} \left(\frac{\lambda}{\eta}\right)^{\lambda} \left(\frac{\lambda}{\eta} + q\right)^{-(N+\lambda)}, \quad (4.59)$$

where  $B(N, \lambda) = \frac{\Gamma(N)\Gamma(\lambda)}{\Gamma(N+\lambda)} = \frac{(N-1)!\Gamma(\lambda)}{\Gamma(N+\lambda)}$ . In this case, we have:

$$\sigma_X^2 = \frac{E_p\{Q_{\bar{\Sigma}}\}}{N} = \frac{\lambda}{\eta(\lambda-1)}, \quad (4.60)$$

$$E_p\{Q_{\bar{\Sigma}}^2\} = \sigma_X^4 \frac{N(N+1)(\lambda-1)}{(\lambda-2)}, \quad \lambda > 2. \quad (4.61)$$

In order to focus on the impact of the mismatch due to the difference between the density generators (or, in other words, in order to make the vector  $\mathbf{r}$  in Eq. (4.18) equals to zero), we assume that  $\sigma_X^2 = \sigma^2$ , so that  $\Sigma_0 = \bar{\Sigma}$ . This guarantees that the MML estimator is *consistent* and  $\lambda$  and  $\eta$  are chosen accordingly. In essence, we can set  $\sigma_X^2 = \sigma^2 = 1$  and then, from Eq. (4.60), we chose  $\lambda$  and  $\eta$  to satisfy  $\eta = \lambda/(\lambda-1)$ . It is worth noting that in practical situations, we have no control on the extra parameters of the true distribution. However, this analysis is useful to better understand the nature of the MCRB and of the MML estimator. The more realistic case in which the power  $\sigma^2$  is jointly estimated with  $\Sigma$  is discussed in Section 4.6.2.

A compact expression for the MCRB for two distributions in the CES family is given in Appendix C. Following the results in Ref. [46] and by applying Eq. (C.10), the MCRB can be expressed as:

$$\text{MCRB}(\bar{\Theta}) = \frac{1}{M} \mathbf{D}_N^\dagger \left[ \frac{1}{(\lambda-2)} \text{vec}(\bar{\Sigma}) \text{vec}(\bar{\Sigma})^T + \frac{(\lambda-1)}{(\lambda-2)} \bar{\Sigma} \otimes \bar{\Sigma} \right] \left( \mathbf{D}_N^\dagger \right)^T. \quad (4.62)$$

where  $\mathbf{D}_N$  is the so-called duplication matrix of order  $N$  [43,47,48]. The duplication matrix is implicitly defined as the unique  $N^2 \times l$  matrix (where  $l = N(N+1)/2$ ) that satisfies the following equality:  $\mathbf{D}_N \text{vecs}(\mathbf{A}) = \text{vec}(\mathbf{A})$  for any symmetric matrix  $\mathbf{A}$ . The symbol  $^\dagger$  denotes the Moore-Penrose pseudo-inverse. Moreover, using the expression of the FIM for  $t$ -distributed data evaluated in Refs. [40] and the properties of the vec and vecs operators, the duplication matrix  $\mathbf{D}_N$  and of the Kronecker product  $\otimes$  [43,47–49], the CRB can be expressed as:

$$\text{CRB}(\bar{\Theta}) = \frac{1}{M} \mathbf{D}_N^\dagger \left[ \frac{N+\lambda+1}{\lambda(N+\lambda)} \text{vec}(\bar{\Sigma}) \text{vec}(\bar{\Sigma})^T + \frac{N+\lambda+1}{N+\lambda} \bar{\Sigma} \otimes \bar{\Sigma} \right] \left( \mathbf{D}_N^\dagger \right)^T, \quad (4.63)$$

For the sake of comparison, in the following figures we report, along with the MSE of the MML, the MCRB and the CRB, as well as the MSE of the robust (unconstrained) Tyler's estimator [50–53]. Tyler's estimator has been derived in the context of the CES distribution as the most robust estimator in min-max sense [53]. In particular, Tyler's estimator can be obtained as the recursive solution of the following (unconstrained) fixed-point (FP) matrix equation:

$$\Sigma = \frac{N}{M} \sum_{m=1}^M \frac{\mathbf{x}_m \mathbf{x}_m^H}{\mathbf{x}_m^H \Sigma^{-1} \mathbf{x}_m}. \quad (4.64)$$

To solve Eq. (4.64), we use the following iterative approach:

$$\begin{cases} \hat{\Sigma}_T^{(0)} = \hat{\Sigma}_{MML} \\ \hat{\Sigma}_T^{(k+1)} = \frac{N}{M} \sum_{m=1}^M \frac{\mathbf{x}_m \mathbf{x}_m^H}{\mathbf{x}_m^H (\hat{\Sigma}_T^{(k)})^{-1} \mathbf{x}_m}, \quad k = 0, \dots, K \end{cases} \quad (4.65)$$

It can be noted that, unlike the recursive procedure proposed in Ref. [50], in Eq. (4.65) there is not a normalization constraint on the trace of  $\hat{\Sigma}_T^{(k)}$ . The MCRB in Eq. (4.9) does not apply to the Tyler's estimator since  $\hat{\Sigma}_T^{(k)}$  has not been derived under any assumed CES distribution.

In order to have a global performance index (i.e., an index that is able to take into account the errors made in the estimation of all the covariance entries), we define  $\varepsilon$  as:

$$\varepsilon \triangleq \left\| E_p \left\{ (\hat{\theta}(\mathbf{x}) - \bar{\theta})(\hat{\theta}(\mathbf{x}) - \bar{\theta})^T \right\} \right\|_F, \quad (4.66)$$

where  $\hat{\theta} = \text{vecs}(\hat{\Sigma})$ ,  $\hat{\Sigma}$  is an estimate of the true covariance matrix  $\Sigma$ ,  $\bar{\theta} = \text{vecs}(\bar{\Sigma})$ , and  $\|\mathbf{A}\|_F = \sqrt{\text{tr}(\mathbf{A}^T \mathbf{A})}$  is the Frobenius norm of matrix  $\mathbf{A}$ . Fig. 4.3 shows the behavior of this global performance index for the MML and Tyler's estimators as a

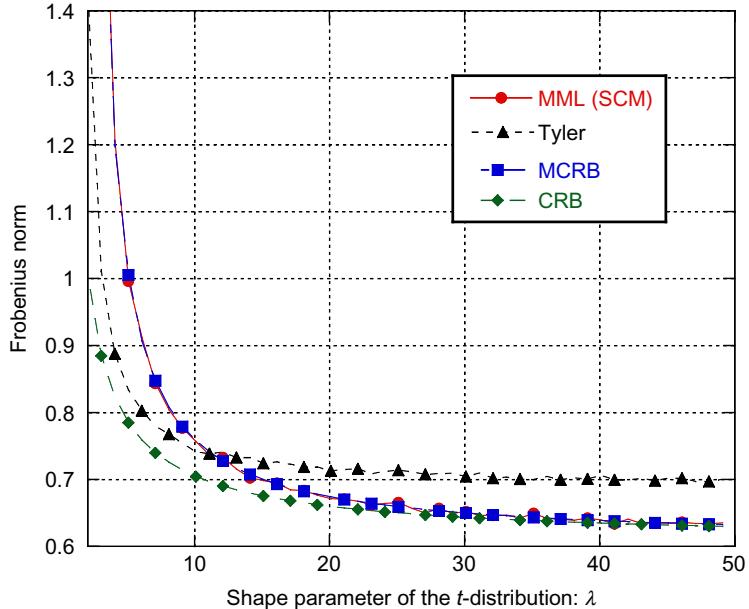


FIG. 4.3

MSE indices,  $\varepsilon_{MML}$  and  $\varepsilon_{Tyler}$ , and bounds  $\varepsilon_{MCRB}$  and  $\varepsilon_{CRB}$  as function of the shape parameter of the  $t$ -distribution ( $\rho=0.8$ ,  $N=8$ ,  $M=3N$ ).

function of the shape parameter  $\lambda$ . As performance bounds, the following quantities are plotted:

$$\varepsilon_{MCRB} \triangleq \|\text{MCRB}(\bar{\theta})\|_F, \quad \varepsilon_{CRB} \triangleq \|\text{CRB}(\bar{\theta})\|_F \quad (4.67)$$

The true covariance matrix is assumed to be  $[\Sigma]_{i,j} = \rho^{|i-j|}$  [54,55]. The value of the one-lag correlation coefficient is  $\rho = 0.8$ , the number of (secondary) data vectors is  $M = 3N$ . To calculate the global performance indices  $\varepsilon_{MML}$  and  $\varepsilon_{Tyler}$  of the estimators, we run  $10^5$  Monte Carlo trials. As expected, for high values of  $\lambda$  the MCRB and the CRB tend to be equal, since for  $\lambda \rightarrow \infty$  the  $t$ -student pdf tends to a complex Gaussian pdf, and the matched and mismatched models tend to coincide. Moreover, as  $\lambda \rightarrow \infty$ , the MML estimator converges to the ML estimator and then it attains the CRB. This is not the case for Tyler's estimator that suffers from "robustness losses," i.e., it is robust but not optimal when the data tends to be Gaussian distributed ( $\lambda \rightarrow \infty$ ). In Fig. 4.4,  $\varepsilon_{MML}$ ,  $\varepsilon_{Tyler}$ ,  $\varepsilon_{MCRB}$ , and  $\varepsilon_{CRB}$  are compared as a function of the number of available data  $M$ , for  $\lambda = 3$ . In this case, Tyler's estimator has better estimation performance than the MML estimator, thanks to its robustness [51,53]. For completeness, in Fig. 4.5 we investigate the performance of the MML and of Tyler's estimator as function of  $\rho$  for  $\lambda = 3$ ,  $N = 8$ , and  $M = 3N$ . As expected, Tyler's estimator achieves better performance than the MML estimator for all the values of  $\rho$ . Finally, it can be noted that the MCRB is not applicable to Tyler's estimator since it is

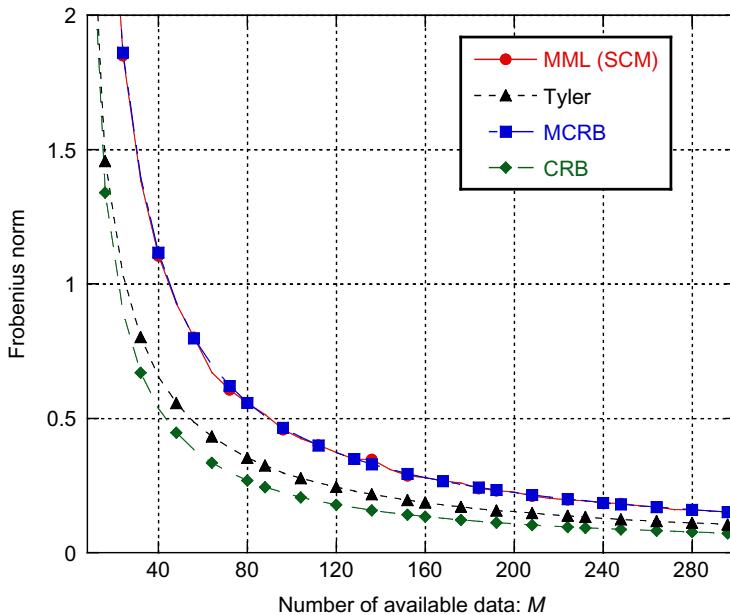
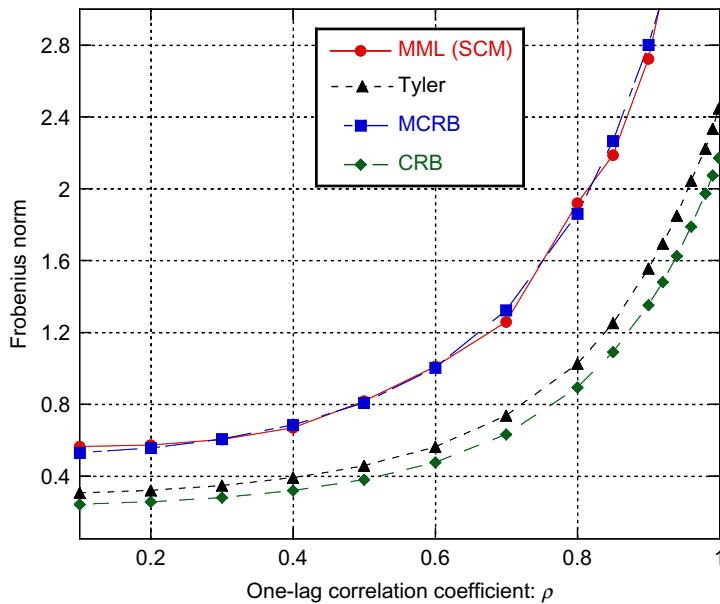


FIG. 4.4

MSE indices,  $\varepsilon_{MML}$  and  $\varepsilon_{Tyler}$ , and bounds  $\varepsilon_{MCRB}$  and  $\varepsilon_{CRB}$  as function of the available data ( $\rho = 0.8$ ,  $N = 8$ ,  $\lambda = 3$ ).

**FIG. 4.5**

MSE indices,  $\varepsilon_{MML}$  and  $\varepsilon_{Tyler}$ , and bounds  $\varepsilon_{MCRB}$  and  $\varepsilon_{CRB}$  as function of the one-lag correlation coefficient  $\rho$  ( $\lambda=3$ ,  $N=8$ ,  $M=3N$ ).

not based on any misspecified data distribution. Therefore its RMSE sometimes falls below the MCRB. On the other hand, since Tyler's estimator is an unbiased estimator of  $\bar{\Sigma}$  (at least in its unconstrained version) its RMSE is always above the CRB.

#### 4.6.1.2 Case Study 2. Assumed pdf: complex Normal, true pdf: Generalized Gaussian

As before, we assume a complex Normal model for the data, while the true data distribution is a GG distribution:

$$p_X(\mathbf{x}_m; \bar{\boldsymbol{\theta}}) \triangleq p_X(\mathbf{x}_m; \bar{\Sigma}) = \frac{\beta \Gamma(N) b^{-N/\beta}}{\pi^N \Gamma(N/\beta)} \frac{1}{|\bar{\Sigma}|} \exp\left(-\frac{(\mathbf{x}_m^H \bar{\Sigma}^{-1} \mathbf{x}_m)^\beta}{b}\right), \quad (4.68)$$

where  $b$  is the scale parameter and  $\beta$  is the shape parameter characterizing the model [21,56]. One advantage of the GG distribution with respect to the  $t$ -distribution is that it can be used to model both the heavy tailed ( $\beta < 1$ ) and the light tailed ( $\beta > 1$ ) distributions as compared to the Normal distribution ( $\beta = 1$ ).

Since we are assuming a complex Normal model for the data, the MML estimator can be derived exactly as discussed in the Case Study 1. In particular, the MML scatter matrix estimator for this mismatched scenario is still the SCM given in Eq. (4.52). Moreover its convergence point, i.e., the point that minimizes the KL divergence

between the true GG distribution and the assumed Normal distribution, is again the matrix  $\Sigma_0 = (\sigma_X^2/\sigma^2)\bar{\Sigma}$  of Eq. (4.57), where  $\sigma_X^2$  is the power related to the GG distribution in Eq. (4.68). As in the Case Study 1, in order to focus our analysis on the effects of the misspecification between the true and the assumed density generators (i.e., in order to make the vector  $r$  in Eq. (4.18) equal to zero), we choose the shape  $\beta$  and scale  $b$  of the GG distribution such that  $\sigma_X^2 = \sigma^2$ , i.e., the MML estimator is consistent. The power  $\sigma_X^2 \triangleq E_p\{Q_{\bar{\Sigma}}\}/N$  of the GG distribution is function of  $\beta$  and  $b$ . In fact, by applying the Stochastic Representation Theorem, it can be shown that  $Q_{\bar{\Sigma}}$  has pdf:

$$p_{Q_{\bar{\Sigma}}}(q) = \frac{\beta q^{N-1}}{b^{N/\beta} \Gamma(N\beta^{-1})} e^{-\frac{q^\beta}{b}}. \quad (4.69)$$

Hence, we have:

$$\sigma_X^2 = E_p\{Q_{\bar{\Sigma}}\}/N = b^{1/\beta} \Gamma\left(\frac{N+1}{\beta}\right)/N \Gamma\left(\frac{N}{\beta}\right), \quad (4.70)$$

$$E_p\{Q_{\bar{\Sigma}}^2\} = b^{2/\beta} \Gamma\left(\frac{N+2}{\beta}\right)/\Gamma\left(\frac{N}{\beta}\right) = \sigma_X^4 N^2 \frac{\Gamma(N/\beta)\Gamma((N+2)/\beta)}{\Gamma((N+1)/\beta)^2}. \quad (4.71)$$

As in the Case Study 1, we set  $\sigma_X^2 = \sigma^2 = 1$ , and then, from Eq. (4.70),  $\beta$  and  $b$  should be chosen to satisfy  $b = [\Gamma(N+1/\beta)/\Gamma(N/\beta)]^\beta$ .

Following the results in Ref. [46] and by applying Eq. (C.10), the MCRB can be expressed as:

$$\text{MCRB}(\bar{\theta}) = \frac{1}{M} \mathbf{D}_N^\dagger \left[ v \bar{\Sigma} \otimes \bar{\Sigma} + (v-1) \text{vec}(\bar{\Sigma}) \text{vec}(\bar{\Sigma})^T \right] \left( \mathbf{D}_N^\dagger \right)^T, \quad (4.72)$$

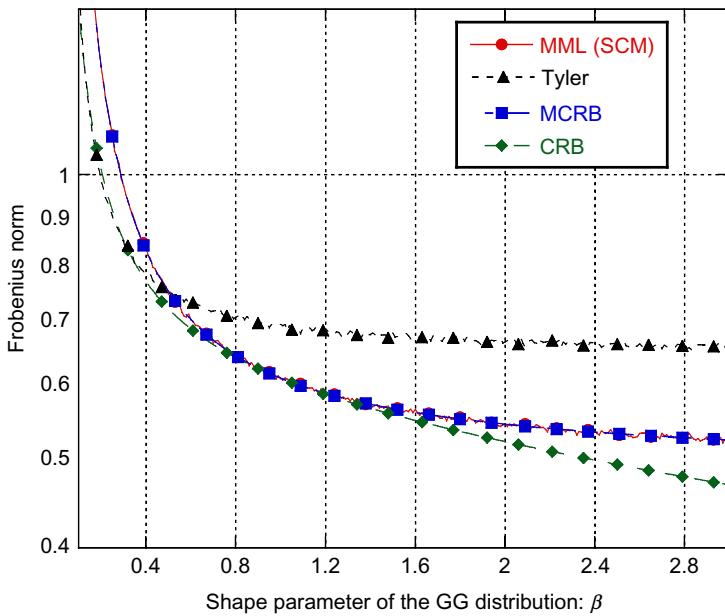
where

$$v \triangleq \frac{N\Gamma(N/\beta)\Gamma((N+2)/\beta)}{(N+1)\Gamma((N+1)/\beta)^2}. \quad (4.73)$$

The FIM for GG-distributed data has been evaluated for a single observation vector  $\mathbf{x}_m$  in Ref. [40]. After some matrix manipulations, the CRB on  $\bar{\Sigma}$  is derived to be:

$$\text{CRB}(\bar{\theta}) = \frac{1}{M} \mathbf{D}_N^\dagger \left[ \frac{N+1}{N+\beta} \bar{\Sigma} \otimes \bar{\Sigma} + \frac{1-\beta}{\beta(N+\beta)} \text{vec}(\bar{\Sigma}) \text{vec}(\bar{\Sigma})^T \right] \left( \mathbf{D}_N^\dagger \right)^T. \quad (4.74)$$

As for the Case Study 1, in Fig. 4.6 we compare the MSE of the MML estimator and of the Tyler's estimator with the MCRB the CRB in terms of the indices  $\epsilon_{MML}$ ,  $\epsilon_{Tyler}$ ,  $\epsilon_{MCRB}$ , and  $\epsilon_{CRB}$ , as a function of the shape parameter  $\beta$ . The value of the one-lag correlation coefficient is  $\rho = 0.8$ , the number of data vectors is  $M = 3N$ . To calculate the MSE of the estimators we run  $10^5$  Monte Carlo trials. As expected, for  $\beta = 1$  the MCRB and the CRB are equal, since the Generalized Gaussian pdf becomes a complex Gaussian pdf, then matched and mismatched cases coincide. In heavy tail disturbance ( $\beta < 1$ ), thanks to its robustness, Tyler's estimator has better performance than the MML estimator. The reverse is true when  $\beta > 1$ .

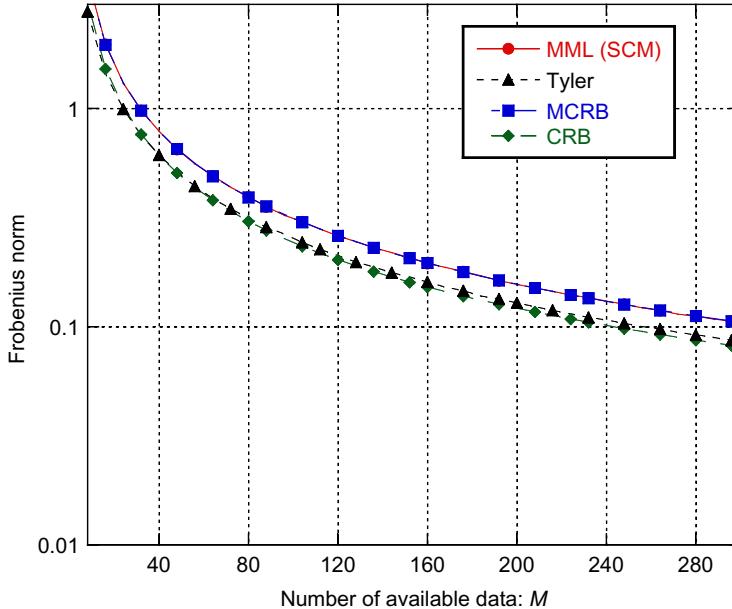
**FIG. 4.6**

MSE indices,  $\varepsilon_{MML}$  and  $\varepsilon_{Tyler}$ , and bounds  $\varepsilon_{MCRB}$  and  $\varepsilon_{CRB}$  as function of the shape parameter of the GG distribution ( $\rho=0.8$ ,  $N=8$ ,  $M=3N$ ).

In Fig. 4.7 the indices  $\varepsilon_{MML}$ ,  $\varepsilon_{Tyler}$ ,  $\varepsilon_{MCRB}$ , and  $\varepsilon_{CRB}$  are compared as function of the number of available data  $M$ . Finally, in Fig. 4.8 the performance of the MML and of the Tyler's estimator are plotted as function of  $\rho$  for  $\beta=0.2$  and  $N=8$ . Even in this case, Tyler's estimator has better performance than the MML estimator and is very close to the CRB.

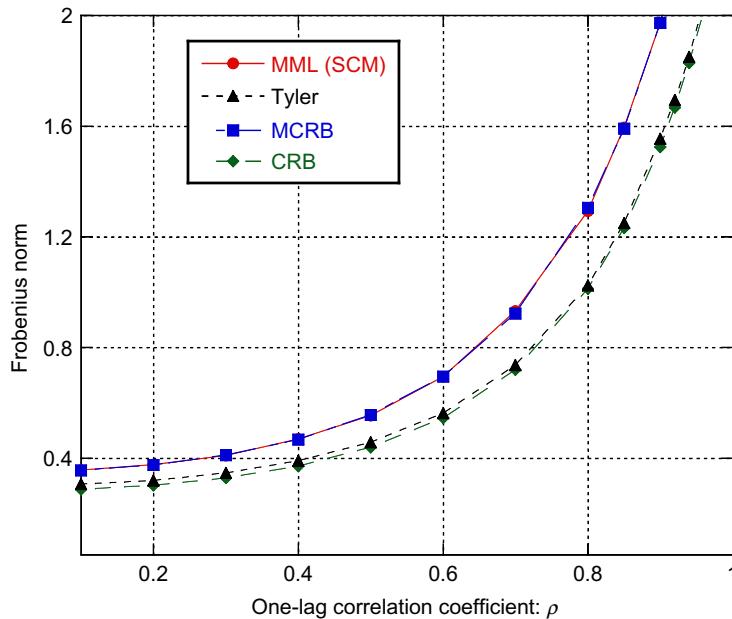
#### 4.6.1.3 Case Study 3. Assumed pdf: Generalized Gaussian; true pdf: t-student

In this study case, we want to investigate the scenario where we know that the data are not Gaussian, but we do not assume the correct “non-Gaussian” model for our data. As in the Case Study 1, we assume that the true distribution is a complex- $t$  distribution, but unlike the previous case, we assume a complex GG distribution for modeling our data. The MML estimator, then, is the ML estimator for the GG data. Unlike the SCM (i.e., the ML estimator for Normal data), the ML estimator for GG data cannot be expressed with an explicit equation but has to be defined through a fixed-point equation. In this subsection, we first discuss some properties of the MML estimator (in particular, bias and consistency in the mismatched sense), and then we evaluate the relevant MCRB. In this case study, the true distribution has the same pdf



**FIG. 4.7**

MSE indices,  $\varepsilon_{MML}$  and  $\varepsilon_{Tyler}$ , and bounds  $\varepsilon_{MCRB}$  and  $\varepsilon_{CRB}$  as function of the available data ( $\rho=0.8$ ,  $N=8$ ,  $\beta=0.2$ ).



**FIG. 4.8**

MSE indices,  $\varepsilon_{MML}$  and  $\varepsilon_{Tyler}$ , and bounds  $\varepsilon_{MCRB}$  and  $\varepsilon_{CRB}$  as function of the one-lag correlation coefficient  $\rho$  ( $\beta=0.2$ ,  $N=8$ ,  $K=3N$ ).

given in Eq. (4.51), while the assumed pdf is the GG distribution in Eq. (4.68) recalled here for sake of clarity:

$$f_X(\mathbf{x}_m; \boldsymbol{\theta}) \triangleq f_X(\mathbf{x}_m; \boldsymbol{\Sigma}) = \frac{\beta \Gamma(N) b^{-N/\beta}}{\pi^N \Gamma(N/\beta)} \frac{1}{|\boldsymbol{\Sigma}|} \exp\left(-\frac{(\mathbf{x}_m^H \boldsymbol{\Sigma}^{-1} \mathbf{x}_m)^\beta}{b}\right),$$

where  $\beta$  is the shape parameter and  $b$  is the scale parameter that are again assumed to be known. In this case, the MML estimator is the solution of the following fixed-point matrix equation [21,40,56]:

$$\hat{\boldsymbol{\Sigma}}_{MML} = \frac{1}{M} \sum_{m=1}^M \varphi\left(\mathbf{x}_m^H \hat{\boldsymbol{\Sigma}}_{MML}^{-1} \mathbf{x}_m\right) \mathbf{x}_m \mathbf{x}_m^H = H_M(\hat{\boldsymbol{\Sigma}}_{MML}), \quad (4.75)$$

where the function  $\varphi$  is given by  $\varphi(t) = (\beta/b)t^{\beta-1}$ . Following Theorem 6 in Ref. [21], it can be shown that, for every (symmetric and positive-definite) starting matrix  $\boldsymbol{\Sigma}^{(0)}$ , the recursive version of Eq. (4.75) converges to  $\hat{\boldsymbol{\Sigma}}_{MML}$ , i.e.,  $\hat{\boldsymbol{\Sigma}}^{k+1} = H_M(\hat{\boldsymbol{\Sigma}}^k) \xrightarrow[k \rightarrow \infty]{} \hat{\boldsymbol{\Sigma}}_{MML}$  iff  $\beta \in (0, 1)$ . For  $\beta > 1$ , i.e., when the tails of the GG distribution are lighter than those of the Normal distribution, the recursive estimator of the scatter matrix is no longer reliable. In fact, when  $\beta > 1$ , the conditions on  $\varphi(t)$  that guarantee the existence and the uniqueness of the estimator are not satisfied. Theorem 5 in Ref. [51] can be used to prove that for  $\beta \in (0, 1)$  we have  $\hat{\boldsymbol{\Sigma}}_{MML}(\mathbf{x}) \xrightarrow[M \rightarrow \infty]{a.s.} \boldsymbol{\Sigma}_0$ , i.e., the MML estimator  $\hat{\boldsymbol{\Sigma}}_{MML}$  converges, almost surely, to  $\boldsymbol{\Sigma}_0$ . From the theory previously discussed, the limiting value  $\boldsymbol{\Sigma}_0$  must be the matrix that minimizes the KL divergence between  $p_X(\mathbf{x}_m)$  and  $f(\mathbf{x}_m; \boldsymbol{\Sigma})$ . In order to calculate  $\boldsymbol{\Sigma}_0$ , we can apply Eq. (4.53), where in this case the density generator is that of the GG distribution, i.e.,  $g(t) = \exp(-t^\beta/b)$ . After some calculations, we get:

$$\partial D(p_X \| f_{\boldsymbol{\Sigma}}) = \text{tr}(\boldsymbol{\Sigma}^{-1} \partial \boldsymbol{\Sigma}) - \frac{\beta}{b} \text{tr}\left(E_p\left\{(\mathbf{x}_m^H \boldsymbol{\Sigma}^{-1} \mathbf{x}_m)^{\beta-1} \boldsymbol{\Sigma}^{-1} \mathbf{x}_m \mathbf{x}_m^H \boldsymbol{\Sigma}^{-1}\right\} \partial \boldsymbol{\Sigma}\right). \quad (4.76)$$

By applying the Stochastic Representation Theorem, we have  $\mathbf{x}_m =_d \sqrt{Q_{\bar{\boldsymbol{\Sigma}}}} \mathbf{T} \mathbf{u}$ , where  $Q_{\bar{\boldsymbol{\Sigma}}} \triangleq \mathbf{z}^H \bar{\boldsymbol{\Sigma}}^{-1} \mathbf{z}$ ,  $\bar{\boldsymbol{\Sigma}} = \mathbf{T} \mathbf{T}^H$  is a factorization of the shape  $\bar{\boldsymbol{\Sigma}}$ ,  $\mathbf{u}$  is a  $N$ -dimensional vector uniformly distributed on the unit hyper-sphere with  $N-1$  topological dimensions such that  $\mathbf{u}^H \mathbf{u} = 1$  and  $E\{\mathbf{u} \mathbf{u}^H\} = N^{-1} \mathbf{I}$ . Then, Eq. (4.76) can be rewritten as:

$$\partial D(p_X \| f_{\boldsymbol{\Sigma}}) = \text{tr}(\boldsymbol{\Sigma}^{-1} \partial \boldsymbol{\Sigma}) - \frac{\beta E\left\{Q_{\bar{\boldsymbol{\Sigma}}}^\beta\right\}}{b} \text{tr}\left(E_p\left\{(\mathbf{u}^H \mathbf{T}^H \boldsymbol{\Sigma}^{-1} \mathbf{T} \mathbf{u})^{\beta-1} \boldsymbol{\Sigma}^{-1} \mathbf{T} \mathbf{u} \mathbf{u}^H \mathbf{T}^H \boldsymbol{\Sigma}^{-1}\right\} \partial \boldsymbol{\Sigma}\right), \quad (4.77)$$

where  $E\left\{Q_{\bar{\boldsymbol{\Sigma}}}^\beta\right\}$  can be evaluated explicitly by using the integral in [57, p. 315, n. 194.3]:

$$E\left\{Q_{\bar{\boldsymbol{\Sigma}}}^\beta\right\} = \left(\frac{\lambda}{\eta}\right)^\beta \frac{\Gamma(\beta+N)\Gamma(\lambda-\beta)}{\Gamma(N)\Gamma(\lambda)}. \quad (4.78)$$

From Eq. (4.77), setting to zero the derivative of the KL divergence w.r.t.  $\Sigma$  leads to:

$$\Sigma^{-1} - \beta b^{-1} E\left\{Q_{\bar{\Sigma}}^{\beta}\right\} E\left\{\left(\mathbf{u}^H \mathbf{T}^H \Sigma^{-1} \mathbf{T} \mathbf{u}\right)^{\beta-1} \Sigma^{-1} \mathbf{T} \mathbf{u} \mathbf{u}^H \mathbf{T}^H \Sigma^{-1}\right\} \Big|_{\Sigma=\Sigma_0} = \mathbf{0}. \quad (4.79)$$

Through some standard matrix manipulation, we get:

$$\mathbf{T}^{-1} \Sigma_0 \mathbf{T}^{-H} = \beta b^{-1} E\left\{Q_{\bar{\Sigma}}^{\beta}\right\} E\left\{\left(\mathbf{u}^H \mathbf{T}^H \Sigma_0^{-1} \mathbf{T} \mathbf{u}\right)^{\beta-1} \mathbf{u} \mathbf{u}^H\right\}. \quad (4.80)$$

Now, assuming that the solution of Eq. (4.80) is a scaled version of the true shape matrix, i.e.,  $\Sigma_0 = \delta \bar{\Sigma}$ , we have  $\delta \mathbf{I} = \frac{\beta}{bN} E\left\{Q_{\bar{\Sigma}}^{\beta}\right\} \delta^{-\beta+1} \mathbf{I}$ , so that

$$\delta = \frac{\lambda}{\eta} \left( \frac{\beta \Gamma(\beta+N) \Gamma(\lambda-\beta)}{\Gamma(N) \Gamma(\lambda)} \right)^{1/\beta}. \quad (4.81)$$

Then, the matrix that minimizes the KL divergence is given by:

$$\Sigma_0 = \frac{\lambda}{\eta} \left( \frac{\beta \Gamma(\beta+N) \Gamma(\lambda-\beta)}{\Gamma(N) \Gamma(\lambda)} \right)^{1/\beta} \bar{\Sigma} \triangleq \delta \bar{\Sigma}. \quad (4.82)$$

Since  $\Sigma_0$  is a scaled version of the true scatter matrix, the MML estimator is not consistent in general. As shown in Refs. [21,51], for the estimator in Eq. (4.75), the following asymptotic relation holds:

$$E_p\left\{\varphi\left(\mathbf{x}_m^H \Sigma_{f,MML}^{-1} \mathbf{x}_m\right) \mathbf{x}_m \mathbf{x}_m^H\right\} = \gamma \bar{\Sigma}. \quad (4.83)$$

Eq. (4.83) can be used to evaluate the bias of the MML estimator in the mismatched sense. The mean value of the MML estimator with respect to the true distribution  $p_X$  is:

$$\boldsymbol{\mu} = E_p\left\{\hat{\Sigma}_{MML}(\mathbf{x})\right\} \underset{M \rightarrow \infty}{=} \gamma \bar{\Sigma}, \quad (4.84)$$

where the scalar term  $\gamma$  can be evaluated by solving the following integral equation [21]:

$$E\left\{\varphi\left(\frac{Q_{\bar{\Sigma}}}{\gamma}\right) \frac{Q_{\bar{\Sigma}}}{\gamma}\right\} = N. \quad (4.85)$$

Given  $\varphi(t) = \beta t^{\beta-1} / b$  and by using the integral in Eq. (4.78),  $\gamma$  can be evaluated as:

$$\gamma = \frac{\lambda}{\eta} \left( \frac{\beta}{bN} \cdot \frac{\Gamma(\beta+N) \Gamma(\lambda-\beta)}{\Gamma(N) \Gamma(\lambda)} \right)^{1/\beta} = \delta. \quad (4.86)$$

Hence, the MML estimator is (asymptotically) *MS-unbiased*, i.e., the mean value  $\boldsymbol{\mu}$  tends to the matrix that minimizes the KL divergence  $\boldsymbol{\mu} = \delta \bar{\Sigma} = \Sigma_0$ . However, the MML is not consistent since it converges to a scaled version of the true scatter matrix. As before, we select the parameter values in such a way that the estimator is consistent and then the vector  $\mathbf{r}$  in Eq. (4.18) is equal to zero. In particular, we

choose a set of shape and scale parameters of the assumed and the true distributions such that  $\delta = 1$ , and then  $\boldsymbol{\mu} = \boldsymbol{\Sigma}_0 = \bar{\boldsymbol{\Sigma}}$ . To have  $\delta = 1$ , a possible choice of the scale parameter  $\eta$  of the  $t$ -distribution and the scale parameter  $b$  of the GG distribution is:

$$\eta = \frac{\lambda}{\lambda - 1} \quad \text{and} \quad b = \frac{\Gamma(\beta + N)\Gamma(\lambda - \beta)}{\Gamma(N)\Gamma(\lambda)} \left( \frac{\lambda}{\eta} \right)^\beta \frac{\beta}{N}.$$

Now, we can compare the estimation performance of the MML estimator directly with the MCRB in Eq. (4.18). As before, the MCRB can be evaluated using the compact expression for  $\mathbf{A}_{\bar{\theta}}^{-1} \mathbf{B}_{\bar{\theta}} \mathbf{A}_{\bar{\theta}}^{-1}$  derived in Appendix C, Eq. (C.10). The density generator for the GG distribution is  $g(t) = \exp(-t^\beta/b)$ , hence we have:

$$\frac{\partial \ln g(Q_\Sigma)}{\partial Q_\Sigma} = -\frac{\beta}{b} Q_\Sigma^{\beta-1} \quad \text{and} \quad \frac{\partial^2 \ln g(Q_\Sigma)}{\partial Q_\Sigma^2} = -\frac{\beta(\beta-1)}{b} Q_\Sigma^{\beta-2}.$$

In order to evaluate the terms  $B_1$ ,  $B_2$ ,  $A_1$ , and  $A_2$  in Eqs. (C.2), (C.3), (C.6), and (C.7), respectively, the integral in Eq. (4.78) is needed. In particular, we have:

$$E \left\{ Q \frac{\partial \ln g(Q)}{\partial Q} \right\} = -\frac{\beta}{b} \left( \frac{\lambda}{\eta} \right)^\beta \frac{\Gamma(\beta + N)\Gamma(\lambda - \beta)}{\Gamma(N)\Gamma(\lambda)}, \quad (4.87)$$

$$E \left\{ Q^2 \left( \frac{\partial \ln g(Q)}{\partial Q} \right)^2 \right\} = \frac{\beta^2}{b^2} \left( \frac{\lambda}{\eta} \right)^{2\beta} \frac{\Gamma(2\beta + N)\Gamma(\lambda - 2\beta)}{\Gamma(N)\Gamma(\lambda)}, \quad (4.88)$$

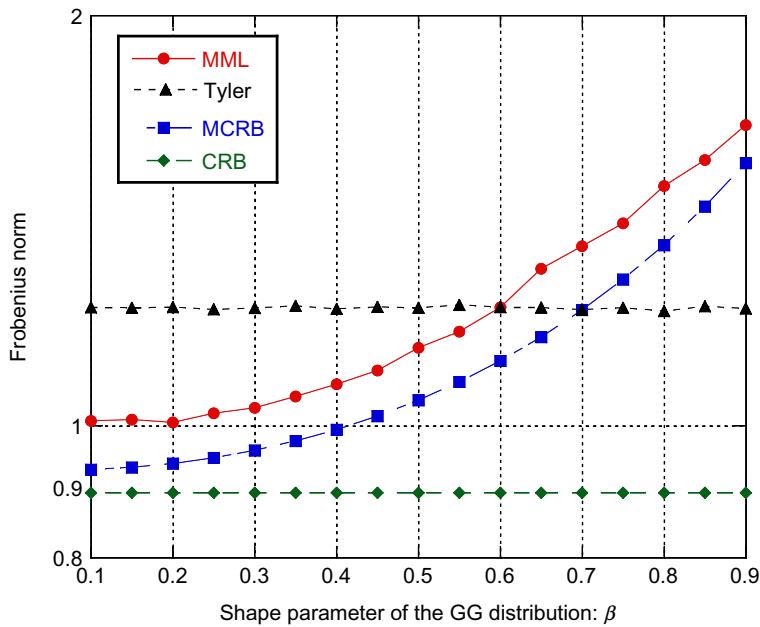
$$E \left\{ Q^2 \frac{\partial^2 \ln g(Q)}{\partial Q^2} \right\} = -\frac{\beta(\beta-1)}{b} \left( \frac{\lambda}{\eta} \right)^\beta \frac{\Gamma(\beta + N)\Gamma(\lambda - \beta)}{\Gamma(N)\Gamma(\lambda)}, \quad (4.89)$$

with  $\beta < \lambda/2$ . Finally, the MCRB is evaluated using Eq. (C.10), while the CRB is given in Eq. (4.63).

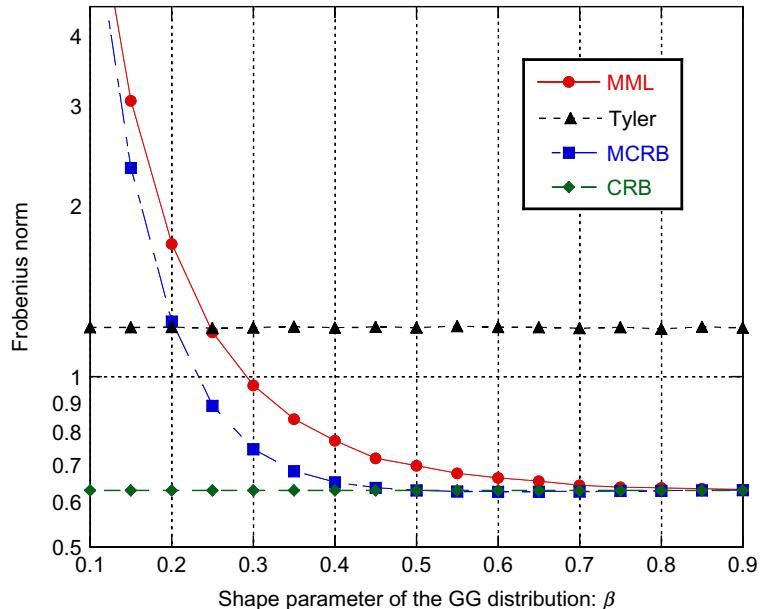
In the following, we compare the MSE of the MML estimator and of Tyler's estimator given in Eq. (4.64) with the MCRB and the CRB by calculating the indices  $\varepsilon_{MML}$ ,  $\varepsilon_{Tyler}$ ,  $\varepsilon_{MCRB}$ , and  $\varepsilon_{CRB}$  previously defined. Both the iterations to derive the MML and Tyler's estimators are initialized using the identity matrix  $\mathbf{I}$ . As before, the value of the one-lag correlation coefficient is  $\rho = 0.8$  and the number of data vectors is  $M = 3N$ . To calculate the MSE of the estimators, we run  $10^5$  Monte Carlo trials.

The simulation results concern two different scenarios: (1) the super-Gaussian scenario, i.e., the true  $t$ -distribution has heavier tails than a Normal distribution and where  $\lambda = 3$ , and (2) the quasi-Gaussian scenario, where  $\lambda = 50$  ( $\lambda$  is the shape parameter of the  $t$ -distribution).

As shown in Fig. 4.9, the MML estimator achieves better performance than Tyler's estimator when  $\beta < 0.6$ , i.e., when the assumed GG distribution has extremely heavy tails. The MCRB gets close to the CRB when  $\beta$  decreases, i.e., the assumed pdf becomes spikier. As expected, in the quasi-Gaussian case of Fig. 4.10, the MML estimator and the MCRB have an opposite behavior with respect

**FIG. 4.9**

MSE indices,  $\varepsilon_{MML}$  and  $\varepsilon_{Tyler}$ , and bounds  $\varepsilon_{MCRB}$  and  $\varepsilon_{CRB}$  as a function of the shape parameter of the GG distribution ( $\rho=0.8$ ,  $N=8$ ,  $M=3N$ ,  $\lambda=3$ ).

**FIG. 4.10**

MSE indices,  $\varepsilon_{MML}$  and  $\varepsilon_{Tyler}$ , and bounds  $\varepsilon_{MCRB}$  and  $\varepsilon_{CRB}$  as a function of the shape parameter of the GG distribution ( $\rho=0.8$ ,  $N=8$ ,  $M=3N$ ,  $\lambda=50$ ).

to the choice of the shape parameter of the assumed GG distribution  $\beta$ . In fact, with  $\lambda=50$ , the true  $t$ -distribution is very close to the Normal distribution, so the performance of the MML estimator becomes better as  $\beta$  tends to 1, i.e., the MML estimator tends to the SCM. Also, the MCRB gets closer and closer to the CRB when  $\beta$  tends to 1. Clearly, the MSE of Tyler's estimator and the CRB do not depend on the value of the shape parameter  $\beta$  of the assumed pdf.

#### 4.6.2 MISSPECIFIED JOINT ESTIMATION OF THE SCATTER MATRIX AND OF THE EXTRA PARAMETERS

In the previous section, we showed how to apply the mismatched estimation framework to the problem of estimating the scatter matrix of a CES distributed random vector when all the extra parameters are *a priori* known. Now, we investigate the more realistic scenario where all these parameters are unknown and should be jointly estimated. In this case, the constraint on the trace of the scatter matrix must be imposed in order to make the joint estimation problem well defined [58].

We investigate the same scenario as in Case Study 1 of Section 4.6.1: the true data pdf is a complex  $t$ -distribution, while the joint MML estimator of the scatter matrix and of the data power is derived under a Normal model assumption. This is a recurring scenario in adaptive radar applications. In fact, the choice of the  $t$ -distribution as true data model has been motivated by experimental evidence (see e.g., [41,42]) that proved its reliability to model spiky clutter data. On the other hand, many radar systems exploit the Normal model for data inference due to its analytical tractability and the consequent real-time implementation of the estimation algorithms based on it.

More formally, we assume that the  $M$  vectors of the available dataset  $\mathbf{x} = \{\mathbf{x}_m\}_{m=1}^M$  are iid and each one has a complex Normal multivariate pdf given in Eq. (4.50):

$$f_X(\mathbf{x}_m; \boldsymbol{\theta}) \triangleq f_X(\mathbf{x}_m; \boldsymbol{\Sigma}, \sigma^2) = \frac{1}{(\pi\sigma^2)^N |\boldsymbol{\Sigma}|} \exp\left(-\frac{\mathbf{x}_m^H \boldsymbol{\Sigma}^{-1} \mathbf{x}_m}{\sigma^2}\right), \quad \text{tr}(\boldsymbol{\Sigma}) = N. \quad (4.90)$$

The covariance matrix is  $\mathbf{M} = E\{\mathbf{x}_m \mathbf{x}_m^H\} = \sigma^2 \boldsymbol{\Sigma}$ , where  $\text{tr}(\boldsymbol{\Sigma}) = N$  and  $\sigma^2$  is the power. Hence, the parameter vector to be estimated can be expressed as  $\boldsymbol{\theta} = [\text{vecs}(\boldsymbol{\Sigma})^T \ \sigma^2]^T \in \Theta$ . However, the true data are distributed according to the complex  $t$ -distribution  $p_X(\mathbf{x}_m; \boldsymbol{\tau}) \triangleq p_X(\mathbf{x}_m; \bar{\boldsymbol{\Sigma}}, \lambda, \eta)$  of Eq. (4.51):

$$p_X(\mathbf{x}_m; \boldsymbol{\tau}) \triangleq p_X(\mathbf{x}_m; \bar{\boldsymbol{\Sigma}}, \lambda, \eta) = \frac{1}{\pi^N |\bar{\boldsymbol{\Sigma}}|} \frac{\Gamma(N+\lambda)}{\Gamma(\lambda)} \left(\frac{\lambda}{\eta}\right)^\lambda \left(\frac{\lambda}{\eta} + \mathbf{x}_m^H \bar{\boldsymbol{\Sigma}}^{-1} \mathbf{x}_m\right)^{-(N+\lambda)}, \quad \text{tr}(\bar{\boldsymbol{\Sigma}}) = N, \quad (4.91)$$

where  $\boldsymbol{\tau} = [\text{vecs}(\bar{\boldsymbol{\Sigma}})^T \ \lambda \ \eta]^T \in \mathcal{T}$  is the true parameter vector and  $\bar{\boldsymbol{\Sigma}}$  is the true scatter matrix that could be different to the scatter matrix  $\boldsymbol{\Sigma}$  of the assumed Gaussian distribution.

It is worth observing that in the mismatched case the parameter space  $\Theta$  that parameterizes the assumed distribution and the parameter space  $T$  that parameterizes the true distribution may be different. In the case at hand, for example,  $T \subset \mathbb{R}^l \times (0, \infty) \times (0, \infty)$  while  $\Theta \subset \mathbb{R}^l \times (0, \infty)$  where  $\times$  indicates the Cartesian product and  $l = N(N+1)/2$  as before. Moreover, the constraint on the trace of the scatter matrix limits both the true and assumed parameter vector to belong to two lower dimensional smooth manifolds  $\bar{T} = \{\tau \in T | \text{tr}(\bar{\Sigma}) = N\}$  and  $\bar{\Theta} = \{\theta \in \Theta | \text{tr}(\Sigma) = N\}$ , respectively.

The main aspects which differentiate the mismatched estimation problem at hand from the one discussed in the Case Study 1 of [Section 4.6.1](#) are as follows:

- The assumed and the true parameter spaces,  $\bar{\Theta}$  and  $\bar{T}$ , are different. Consequently, the simplified approach discussed in [Section 4.4.4](#) cannot be applied and a consistent MML estimator does not exist.
- No assumptions are made on the value of  $\sigma^2$ , i.e., the power of the assumed Normal distribution. It is considered an additional unknown parameter that needs to be jointly estimated with the scatter matrix  $\Sigma$ .
- To guarantee the identifiability of  $\sigma^2$  and  $\Sigma$ , a constraint on  $\Sigma$ , i.e.,  $\text{tr}(\Sigma) = N$ , needs to be imposed. This means that we should compare the performance of a constrained MML estimator with the CMCRB derived in Eq. (4.27).

In the following, the constrained MML estimator for the estimation of  $\theta$  is first derived and its convergence properties investigated. Then, the CMCRB for the joint estimation problem at hand is calculated.

#### 4.6.2.1 Derivation of the constrained MML (CMMML) estimator

In order to obtain an estimation of  $\theta$ , we apply the MML algorithm in Eq. (4.13). In particular, under the assumption of complex Normal data model in Eq. (4.90), the likelihood function can be expressed as:

$$L(\theta) = \sum_{m=1}^M \ln f_X(\mathbf{x}_m; \theta) = -NM \ln \sigma^2 - M \ln |\Sigma| - \sum_{m=1}^M \mathbf{x}_m^H \Sigma^{-1} \mathbf{x}_m / \sigma^2. \quad (4.92)$$

Then, the MML estimator can be obtained by maximizing  $L(\theta)$  subject to the linear constraint  $\text{tr}(\Sigma) = N$ . To proceed, we do not rely on the Lagrange multiplier method, but rather we follow a different, yet equivalent (at least asymptotically), procedure [59]. We first derive the unconstrained MML estimator and then we project it on the lower dimensional manifold  $\bar{\Theta}$  by imposing the constraint. Specifically, the MML estimator is the solution of the following problem:

$$\begin{cases} \frac{\partial L(\theta)}{\partial \sigma^2} = -\frac{NM}{\sigma^2} + \frac{1}{\sigma^4} \sum_{m=1}^M \mathbf{x}_m^H \Sigma^{-1} \mathbf{x}_m = 0 \\ \frac{\partial L(\theta)}{\partial \Sigma} = -M\Sigma^{-1} + \frac{\Sigma^{-1}}{\sigma^2} \sum_{m=1}^M \mathbf{x}_m \mathbf{x}_m^H \Sigma^{-1} = 0 \\ \text{s.t. } \text{tr}(\Sigma) = N \end{cases} \quad (4.93)$$

Then, we have:

$$\begin{cases} \hat{\sigma}_{MML}^2 = \frac{1}{NM} \sum_{m=1}^M \mathbf{x}_m^H \hat{\Sigma}_{MML}^{-1} \mathbf{x}_m \\ \hat{\Sigma}_{MML} = \frac{N}{\sum_{m=1}^M \mathbf{x}_m^H \hat{\Sigma}_{MML}^{-1} \mathbf{x}_m} \sum_{m=1}^M \mathbf{x}_m \mathbf{x}_m^H \\ \text{s.t. } \text{tr}(\hat{\Sigma}_{MML}) = N \end{cases} \quad (4.94)$$

Hence, imposing the constraint we get the CMMML estimators of  $\sigma^2$  and  $\Sigma$ :

$$\begin{cases} \hat{\sigma}_{CMMML}^2 = \frac{1}{NM} \sum_{m=1}^M \mathbf{x}_m^H \hat{\Sigma}_{CMMML}^{-1} \mathbf{x}_m \\ \hat{\Sigma}_{CMMML} = \frac{N}{\sum_{m=1}^M \mathbf{x}_m^H \mathbf{x}_m} \sum_{m=1}^M \mathbf{x}_m \mathbf{x}_m^H \end{cases} \quad (4.95)$$

Now we need to find the convergence point  $\boldsymbol{\theta}_0$  of the CMMML estimator (see Eq. 4.14). As discussed in Section 4.4.3, the convergence point  $\boldsymbol{\theta}_0 = [\text{vecs}(\boldsymbol{\Sigma}_0)^T \ \sigma_0^2]^T$  is the vector that minimizes the KL divergence between the true  $t$ -distribution  $p_X(\mathbf{x}_m; \boldsymbol{\tau})$  in Eq. (4.91) and  $f_X(\mathbf{x}_m; \boldsymbol{\theta})$  in Eq. (4.90). To this end, we have to solve the following system:

$$\begin{cases} \frac{\partial D(p_X \| f_{\boldsymbol{\theta}})}{\partial \sigma^2} = -E_p \left\{ \frac{\partial \ln f_X(\mathbf{x}_m; \boldsymbol{\Sigma}, \sigma^2)}{\partial \sigma^2} \right\} = 0 \\ \frac{\partial D(p_X \| f_{\boldsymbol{\theta}})}{\partial \boldsymbol{\Sigma}} = -E_p \left\{ \frac{\partial \ln f_X(\mathbf{x}_m; \boldsymbol{\Sigma}, \sigma^2)}{\partial \boldsymbol{\Sigma}} \right\} = \mathbf{0} \end{cases} \quad (4.96)$$

The first equation immediately provides:

$$\frac{\partial D(p_X \| f_{\boldsymbol{\theta}})}{\partial \sigma^2} = E_p \left\{ \frac{\partial}{\partial \sigma^2} \left( N \ln \sigma^2 + \frac{\mathbf{x}_m^H \boldsymbol{\Sigma}^{-1} \mathbf{x}_m}{\sigma^2} \right) \right\} = \frac{N}{\sigma^2} - \frac{E_p\{Q_{\boldsymbol{\Sigma}}\}}{\sigma^4} = 0, \quad (4.97)$$

where  $Q_{\boldsymbol{\Sigma}} \triangleq \mathbf{x}_m^H \boldsymbol{\Sigma}^{-1} \mathbf{x}_m$ . By solving Eq. (4.97) we get:  $\sigma_0^2 = \frac{E_p\{Q_{\boldsymbol{\Sigma}}\}}{N}$ .

The derivative of the KL divergence with respect to  $\boldsymbol{\Sigma}$  is instead given by:

$$\frac{\partial D(p_X \| f_{\boldsymbol{\theta}})}{\partial \boldsymbol{\Sigma}} = \boldsymbol{\Sigma}^{-1} - \frac{E\{Q_{\boldsymbol{\Sigma}}\}}{N\sigma^2} \boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma} \boldsymbol{\Sigma}^{-1} = \mathbf{0}, \quad \text{tr}(\boldsymbol{\Sigma}) = \text{tr}(\bar{\boldsymbol{\Sigma}}) = N, \quad (4.98)$$

whose solution is  $\boldsymbol{\Sigma}_0 = \frac{E\{Q_{\boldsymbol{\Sigma}_0}\}}{N\sigma^2} \bar{\boldsymbol{\Sigma}}$ . Putting together the two solutions, we finally obtain:

$$\sigma_0^2 = \frac{E\{Q_{\boldsymbol{\Sigma}_0}\}}{N} \quad \text{and} \quad \boldsymbol{\Sigma}_0 = \frac{E\{Q_{\boldsymbol{\Sigma}_0}\}}{N\sigma_0^2} \bar{\boldsymbol{\Sigma}} = \bar{\boldsymbol{\Sigma}}, \quad \text{where } \text{tr}(\boldsymbol{\Sigma}_0) = \text{tr}(\bar{\boldsymbol{\Sigma}}) = N, \quad (4.99)$$

and

$$\sigma_0^2 = \frac{E\{Q_{\boldsymbol{\Sigma}_0}\}}{N} = \frac{E\{Q_{\bar{\boldsymbol{\Sigma}}}\}}{N} = \frac{E_p \left\{ \mathbf{x}_m^H \bar{\boldsymbol{\Sigma}}^{-1} \mathbf{x}_m \right\}}{N} = \bar{\sigma}^2 = \frac{\lambda}{\eta(\lambda-1)}, \quad (4.100)$$

where  $\bar{\sigma}^2$  is the true statistical power of the data. Eqs. (4.99)–(4.100) show that the CMMML estimator converges a.s. to the parameter vector  $\hat{\theta}_{CMMML}(\mathbf{x}) \xrightarrow[M \rightarrow \infty]{a.s.} \boldsymbol{\theta}_0 = \left[ \text{vecs}(\bar{\Sigma})^T \quad \bar{\sigma}^2 \right]^T$ , i.e.,

$$\hat{\sigma}_{CMMML}^2(\mathbf{x}) \xrightarrow[M \rightarrow \infty]{a.s.} \bar{\sigma}^2 = \lambda/\eta(\lambda - 1) \quad \text{and} \quad \hat{\Sigma}_{CMMML}(\mathbf{x}) \xrightarrow[M \rightarrow \infty]{a.s.} \bar{\Sigma}. \quad (4.101)$$

Hence, it provides “consistent” estimates for both the scatter matrix and the power of the true data model [58,61]. From a practical point of view, this means that we can use the simple mismatched estimator based on the Gaussian assumption to estimate the scatter matrix and the average power of the (complex  $t$ -distributed) data since it converges to the true required quantities. The analysis of the performance loss of the mismatched estimator in Eq. (4.95) is reported in the next subsection (see Ref. [53] for more details).

#### 4.6.2.2 The CMCRB for the joint estimation of the scatter matrix and the power

In Section 4.6.1 (Case Study 1), the MCRB on the estimation of the scatter matrix has been evaluated for a complex- $t$  distribution when the assumed misspecified distribution is a complex Normal pdf, under the assumption of a priori known power. Here, we generalize the result for the case of unknown power, i.e., when the power  $\sigma^2$  and the scatter matrix  $\Sigma$  are unknown and jointly estimated. In this case  $\sigma^2$  and  $\Sigma$  are not identifiable unless a constraint on  $\Sigma$  is imposed, e.g.,  $\text{tr}(\Sigma) = N$ . In order to incorporate this constraint in the MCRB, we calculate the CMCRB of Theorem 4.3 in Section 4.4.5. In the following, we specialize the general expression provided in Eq. (4.27) for the case study at hand.

##### Evaluation of the matrix $\mathbf{A}_{\boldsymbol{\theta}_0}$

Matrix  $\mathbf{A}_{\boldsymbol{\theta}_0}$  can be decomposed in the following blocks:

$$\mathbf{A}_{\boldsymbol{\theta}_0} = \mathbf{T}_1^T \begin{bmatrix} \mathbf{A}_{\bar{\Sigma}} & \mathbf{A}_c \\ \mathbf{A}_c^T & A_{\bar{\sigma}^2} \end{bmatrix} \mathbf{T}_1, \quad (4.102)$$

$$\mathbf{T}_i = \begin{bmatrix} \mathbf{D}_N & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_i \end{bmatrix}, \quad (4.103)$$

where  $\mathbf{I}_i$  is the identity matrix of dimension  $i \times i$  and  $\mathbf{D}_N$  is the so-called duplication matrix of order  $N$  [43]. Following the procedure in Ref. [46], we have:

$$\mathbf{A}_{\bar{\Sigma}} = -\bar{\Sigma}^{-1} \otimes \bar{\Sigma}^{-1}. \quad (4.104)$$

$$\begin{aligned}
 A_{\bar{\sigma}^2} &= E_p \left\{ \frac{\partial^2}{\partial^2 \bar{\sigma}^2} \ln f_X(\mathbf{x}_m; \boldsymbol{\theta}_0) \right\} = E_p \left\{ \frac{\partial^2}{\partial^2 \bar{\sigma}^2} \left( -N \ln \bar{\sigma}^2 - \frac{Q_{\bar{\Sigma}}}{\bar{\sigma}^2} \right) \right\} \\
 &= \frac{N}{\bar{\sigma}^4} - 2 \frac{E_p \{ Q_{\bar{\Sigma}} \}}{\bar{\sigma}^6} = -\frac{N}{\bar{\sigma}^4},
 \end{aligned} \tag{4.105}$$

$$\begin{aligned}
 [\mathbf{A}_c]_{i,1} &= E_p \left\{ \frac{\partial^2 \ln f_X(\mathbf{x}_m; \boldsymbol{\theta}_0)}{\partial \bar{\sigma}^2 \partial \text{vec}(\bar{\Sigma})_i} \right\} = -\frac{1}{\bar{\sigma}^4} E_p \left\{ \mathbf{x}_m^H \bar{\Sigma}^{-1} \mathbf{A}_i \bar{\Sigma}^{-1} \mathbf{x}_m \right\} \\
 &= -\frac{1}{\bar{\sigma}^2} \text{tr}(\bar{\Sigma}^{-1} \mathbf{A}_i) = -\frac{1}{\bar{\sigma}^2} \text{vec}(\bar{\Sigma}^{-1})^T \text{vec}(\mathbf{A}_i)
 \end{aligned} \tag{4.106}$$

where  $\mathbf{A}_i = \partial \bar{\Sigma} / \partial \theta_i$  is a symmetric 0-1 matrix. From Eq. (4.106) we get:  
 $\mathbf{A}_c = -\frac{1}{\bar{\sigma}^2} \text{vec}(\bar{\Sigma}^{-1})$ .

### Evaluation of the matrix $\mathbf{B}_{\boldsymbol{\theta}_0}$

Matrix  $\mathbf{B}_{\boldsymbol{\theta}_0}$  can be decomposed in the following blocks:

$$\mathbf{B}_{\boldsymbol{\theta}_0} = \mathbf{T}_1^T \begin{bmatrix} \mathbf{B}_{\bar{\Sigma}} & \mathbf{B}_c \\ \mathbf{B}_c^T & B_{\bar{\sigma}^2} \end{bmatrix} \mathbf{T}_1. \tag{4.107}$$

As before, following the procedure in Ref. [46], we get:

$$\mathbf{B}_{\bar{\Sigma}} = \frac{1}{\lambda - 2} \text{vec}(\bar{\Sigma}^{-1}) \text{vec}(\bar{\Sigma}^{-1})^T + \frac{\lambda - 1}{\lambda - 2} \bar{\Sigma}^{-1} \otimes \bar{\Sigma}^{-1}. \tag{4.108}$$

$$\begin{aligned}
 B_{\bar{\sigma}^2} &= E_p \left\{ \left[ \frac{\partial \ln f_X(\mathbf{x}_m; \boldsymbol{\theta}_0)}{\partial \bar{\sigma}^2} \right]^2 \right\} = E_p \left\{ \left[ \frac{\partial}{\partial \bar{\sigma}^2} \left( -N \ln \bar{\sigma}^2 - \frac{Q_{\bar{\Sigma}}}{\bar{\sigma}^2} \right) \right]^2 \right\} \\
 &= \frac{N^2}{\bar{\sigma}^4} - 2 \frac{E_p \{ Q_{\bar{\Sigma}} \}}{\bar{\sigma}^6} + \frac{E_p \{ Q_{\bar{\Sigma}}^2 \}}{\bar{\sigma}^8} = \frac{N^2}{\bar{\sigma}^4} - 2 \frac{N^2}{\bar{\sigma}^4} + \frac{N(N+1)(\lambda-1)}{\bar{\sigma}^4(\lambda-2)} \\
 &= \frac{N(N+\lambda-1)}{\bar{\sigma}^4(\lambda-2)}.
 \end{aligned} \tag{4.109}$$

$$\begin{aligned}
 [\mathbf{B}_c]_{i,1} &= E_p \left\{ \frac{\partial \ln f_X(\mathbf{x}_m; \boldsymbol{\theta}_0)}{\partial \bar{\sigma}^2} \cdot \frac{\partial \ln f_X(\mathbf{x}_m; \boldsymbol{\theta}_0)}{\partial \text{vec}(\bar{\Sigma})_i} \right\} \\
 &= E_p \left\{ \left( \frac{E_p \{ Q_{\bar{\Sigma}} \}}{\bar{\sigma}^4} - \frac{N}{\bar{\sigma}^2} \right) \left( \frac{1}{\bar{\sigma}^2} \mathbf{x}_m^H \bar{\Sigma}^{-1} \mathbf{A}_i \bar{\Sigma}^{-1} \mathbf{x}_m - \text{tr}(\bar{\Sigma}^{-1} \mathbf{A}_i) \right) \right\} \\
 &= -\frac{N}{\bar{\sigma}^2} \text{tr}(\bar{\Sigma}^{-1} \mathbf{A}_i) + \frac{1}{\bar{\sigma}^6} E_p \left\{ \mathbf{x}_m^H \bar{\Sigma}^{-1} \mathbf{x}_m \mathbf{x}_m^H \bar{\Sigma}^{-1} \mathbf{A}_i \bar{\Sigma}^{-1} \mathbf{x}_m \right\} \\
 &= \left( -\frac{N}{\bar{\sigma}^2} + \frac{(N+1)(\lambda-1)}{\bar{\sigma}^2(\lambda-2)} \right) \text{tr}(\bar{\Sigma}^{-1} \mathbf{A}_i) = \frac{N+\lambda-1}{\bar{\sigma}^2(\lambda-2)} \text{tr}(\bar{\Sigma}^{-1} \mathbf{A}_i).
 \end{aligned} \tag{4.110}$$

Hence, we get  $\mathbf{B}_c = \frac{N+\lambda-1}{\bar{\sigma}^2(\lambda-2)} \text{vec}(\bar{\Sigma}^{-1})$ .

### Evaluation of the matrix $\mathbf{U}$

The continuously differentiable constraint  $\text{tr}(\boldsymbol{\Sigma}) = N$  can be rewritten as:

$$f(\boldsymbol{\theta}) = \sum_{i \in I} \text{vecs}(\boldsymbol{\Sigma})_i - N = 0, \quad (4.111)$$

where  $I$  is the set of the indices of the diagonal entries of  $\boldsymbol{\Sigma}$  that can be explicitly described as:

$$I = \left\{ i \mid i = 1 + N(j-1) - \frac{(j-1)(j-2)}{2}, j = 1, \dots, N \right\}. \quad (4.112)$$

Following Ref. [34], we define the  $(l+1)$ -dimensional gradient vector as:

$$\nabla f(\boldsymbol{\theta}) = \frac{\partial f(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^T} = \begin{bmatrix} \frac{\partial}{\partial \text{vecs}(\boldsymbol{\Sigma})^T} \sum_{i \in I} \text{vecs}(\boldsymbol{\Sigma})_i & 0 \end{bmatrix} = [\mathbf{1}_l^T \ 0]^T, \quad (4.113)$$

where  $\mathbf{1}_l$  is a  $l$ -dimensional column vector defined as:

$$[\mathbf{1}_l]_i = \begin{cases} 1 & i \in I \\ 0 & \text{otherwise} \end{cases}. \quad (4.114)$$

The gradient  $\nabla f(\boldsymbol{\theta})$  has clearly full row rank and hence there exists a matrix  $\mathbf{U} \in \mathbb{R}^{(l+1) \times l}$  whose columns form an orthonormal basis for the null space of  $\nabla f(\boldsymbol{\theta})$ , that is,  $\nabla f(\boldsymbol{\theta})\mathbf{U} = \mathbf{0}$  where  $\mathbf{U}^T\mathbf{U} = \mathbf{I}$ . The matrix  $\mathbf{U}$  can be obtained numerically by evaluating, e.g., using the singular value decomposition (SVD), the  $l$  orthonormal eigenvectors associated to the zero eigenvalue of  $\nabla f(\boldsymbol{\theta})$  in Eq. (4.113).

#### 4.6.2.3 Performance analysis

We now compare the estimation performance of the CMMML estimator of Eq. (4.95) [60] with the CMCRB. For the sake of comparison, in the following figures, we also report the MSE of the constrained Tyler's estimator (C-Tyler) and the (matched) CCRB for the joint estimation of  $\boldsymbol{\tau} = [\text{vecs}(\boldsymbol{\Sigma})^T \ \lambda \ \eta]^T$  derived in Ref. [61,62]. In particular, the C-Tyler estimator is the constrained version of the robust estimator discussed in Eq. (4.64) that can be evaluated by using the following iterative approach proposed in Ref. [50]:

$$\begin{cases} \hat{\boldsymbol{\Sigma}}_T^{(0)} = \mathbf{I} \\ \mathbf{S}_T^{(k+1)} = \sum_{m=1}^M \frac{\mathbf{x}_m \mathbf{x}_m^H}{\mathbf{x}_m^H (\hat{\boldsymbol{\Sigma}}_T^{(k)})^{-1} \mathbf{x}_m} \\ \hat{\boldsymbol{\Sigma}}_T^{(k+1)} = N \mathbf{S}_T^{(k+1)} / \text{tr}(\mathbf{S}_T^{(k+1)}) \end{cases} \quad (4.115)$$

for  $k = 1, \dots, K$ , where  $K$  is the number of iterations. It can be noted that in Eq. (4.115) there is a normalization on the trace of  $\hat{\boldsymbol{\Sigma}}_T^{(k)}$  at every step of the iterative procedure to impose the constraint on the trace. Asymptotic consistency and unbiasedness properties are discussed in Ref. [21,53]. It is worth noting that the performance of the C-Tyler estimator can be assessed by comparing its error covariance matrix on the estimation of  $\boldsymbol{\Sigma}$  with the CCRB derived in Ref. [61,62].

As global MSE index for the scatter matrix estimators, we use the one defined in Eq. (4.66) (e.g.,  $\varepsilon_{CMML}$  and  $\varepsilon_{C-Tyler}$ ), while as performance bounds, the following quantity is plotted:

$$\varepsilon_{CMCRB} \triangleq \|CMCRB(\boldsymbol{\Sigma})\|_F, \quad \varepsilon_{CCRB} \triangleq \|CCRB(\boldsymbol{\Sigma})\|_F. \quad (4.116)$$

The accuracy on the estimate of average power  $\sigma^2$  in the mismatched case is measured through its MSE, which is compared with the CMCRB. To calculate the estimation accuracy, we run  $10^5$  Monte Carlo trials. The simulation results have been organized as follows:

1. Estimation accuracy as function of the number  $M$  of available data vectors (Figs. 4.11 and 4.12). Simulation parameters:  $\rho=0.8, N=16, \lambda=3, \eta=1, K=4$ .
2. Estimation accuracy as function of the shape parameter  $\lambda$  (Figs. 4.13 and 4.14). Simulation parameters:  $\rho=0.8, N=16, M=10N, \eta=1, K=4$ .
3. Estimation accuracy as function of the one-lag correlation coefficient  $\rho$  (Figs. 4.15 and 4.16). Simulation parameters:  $N=16, M=10N, \lambda=3, \eta=1, K=4$ .

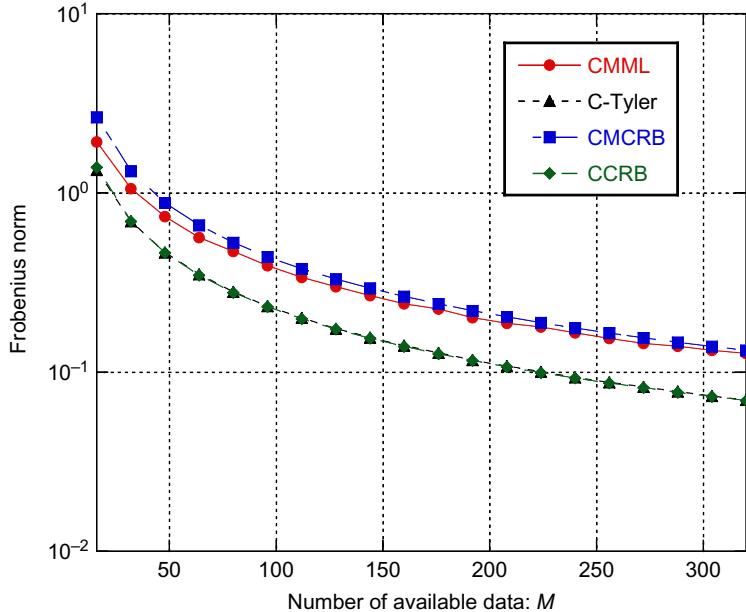
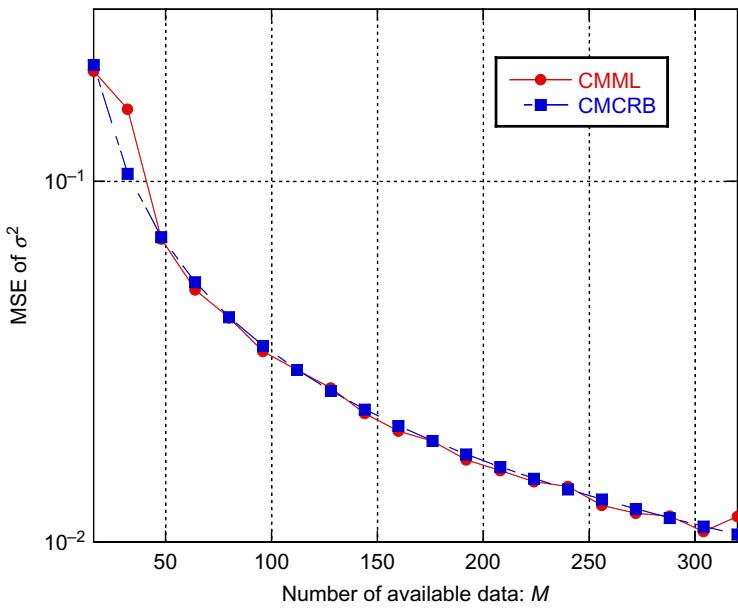


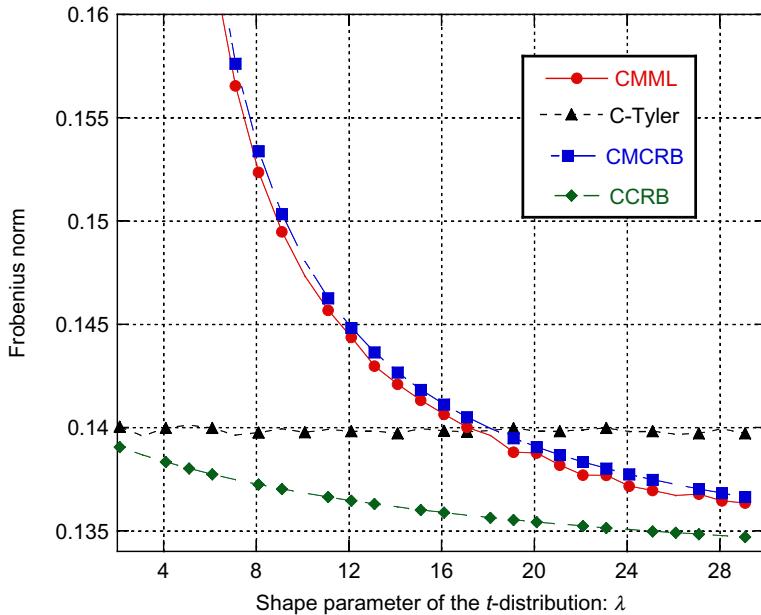
FIG. 4.11

MSE indices,  $\varepsilon_{CMML}$  and  $\varepsilon_{C-Tyler}$ , and bounds  $\varepsilon_{CMCRB}$  and  $\varepsilon_{CCRB}$  as function of the number  $M$  of available data vectors ( $\rho=0.8, N=16, \lambda=3, \eta=1$ ).



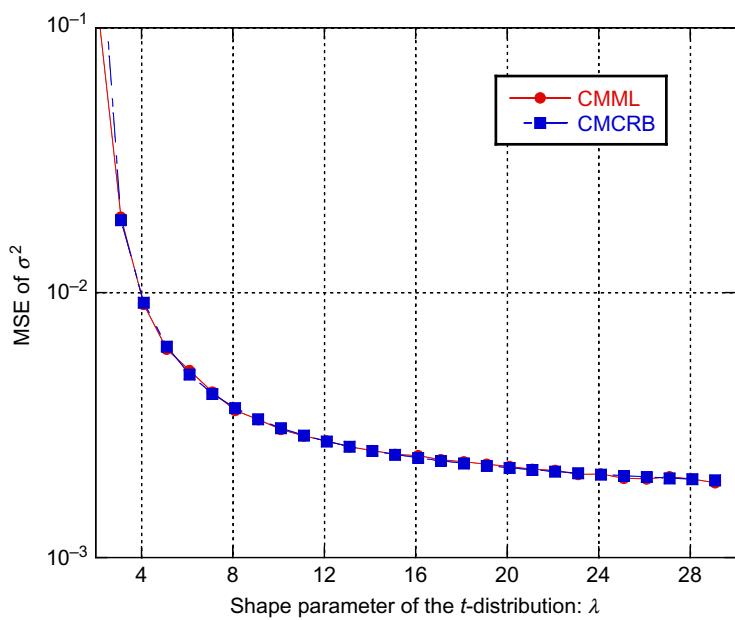
**FIG. 4.12**

MSE of the CMMML estimator of  $\sigma^2$  and CMCRB as function of the number  $M$  of available data vectors ( $\rho=0.8$ ,  $N=16$ ,  $\lambda=3$ ,  $\eta=1$ ).

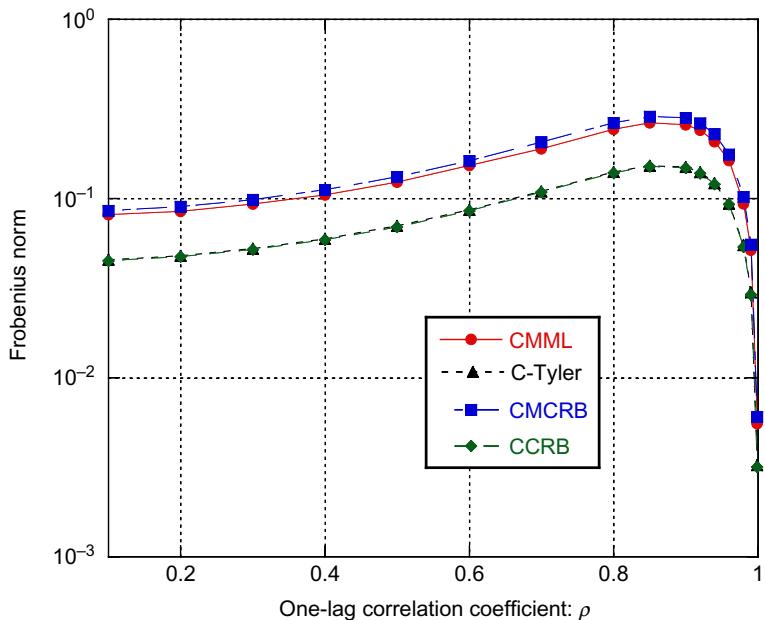


**FIG. 4.13**

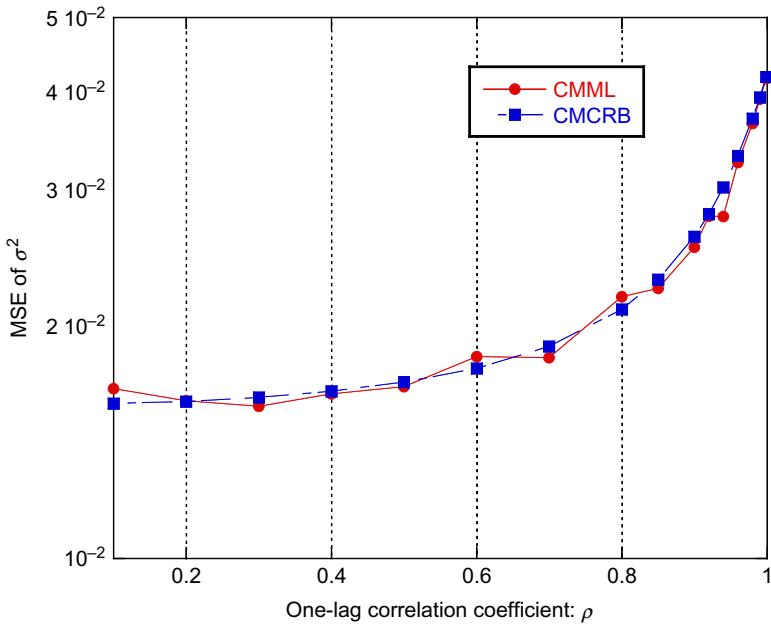
The MSE indices  $\varepsilon_{CMMML}$  and  $\varepsilon_{C\text{-}Tyler}$ , and bounds  $\varepsilon_{CMCRB}$  and  $\varepsilon_{CCRB}$  as function of the shape parameter  $\lambda$  ( $\rho=0.8$ ,  $N=16$ ,  $M=10N$ ,  $\eta=1$ ).

**FIG. 4.14**

MSE of the CMMLE estimator of  $\sigma^2$  and CMCRB as function of the shape parameter  $\lambda$  ( $\rho=0.8$ ,  $N=16$ ,  $M=10N$ ,  $\eta=1$ ).

**FIG. 4.15**

MSE indices  $\varepsilon_{CMMLE}$  and  $\varepsilon_{C-Tyler}$ , and bounds  $\varepsilon_{CMCRB}$  and  $\varepsilon_{CCRB}$  as function of the one-lag correlation coefficient  $\rho$  ( $N=16$ ,  $M=10N$ ,  $\lambda=3$ ,  $\eta=1$ ).

**FIG. 4.16**

MSE of the CMML estimator of  $\sigma^2$  and CMCRB as function of the one-lag correlation coefficient  $\rho$  ( $N=16$ ,  $M=10N$ ,  $\lambda=3$ ,  $\eta=1$ ).

Based on the numerical analysis, we observe that:

- Regarding the CMML estimator, it always achieves the CMCRB, both for the estimation of the scatter matrix and for the estimation of the average power. The CMML presents a small bias on the estimation of the scatter matrix. Hence,  $\hat{\Sigma}_{CMML}$  is not a MS-unbiased estimator (at least in the finite sample regime) [58]. For this reason,  $\epsilon_{CMML}$  can be slightly below the CMCRB. The loss in estimation accuracy due to the mismatch is particularly high for extremely heavy tailed data, i.e., when  $\lambda$  is small (see Fig. 4.13). When  $\lambda \rightarrow 0$ , the CMCRB rapidly increases while the CCRB is quite independent of  $\lambda$ . On the other hand, when  $\lambda \rightarrow \infty$ , the CMCRB and the CCRB tend to coincide, as expected.
- Regarding the scatter matrix estimation, the robust C-Tyler estimator is an “almost” efficient estimator, even if it is not the most efficient estimator for  $t$ -distributed data. The MSE index  $\epsilon_{C-Tyler}$  is close to the CCRB especially for small  $\lambda$  (see Figs. 4.11, 4.13, and 4.15). In particular, its performance is robust, i.e., it is not affected by the value of the shape parameter  $\lambda$  (see Fig. 4.13), even if it is not efficient for large  $\lambda$ .

## 4.7 HYPOTHESIS TESTING PROBLEM FOR TARGET DETECTION

The last section of this chapter focuses on the target detection problem. In particular, this section aims at comparing the detection performance of the *adaptive normalized matched filter* (ANMF) by exploiting the CMML and the C-Tyler estimators for the scatter matrix. The normalized matched filter (NMF) has been derived and analyzed by many authors under different names (see, e.g., [54,55,63–70]) in its adaptive and nonadaptive (i.e., when the disturbance scatter matrix is assumed to be known) versions. One of the most remarkable properties of the nonadaptive NMF is the fact that it is a distribution-free detector under CES distributed clutter, i.e., the pdf of the decision statistic is invariant w.r.t. the particular CES distribution followed by the clutter [21]. We now summarize briefly the classical radar detection problem.

The problem is to detect the possible presence of a complex signal vector  $\mathbf{s}$  in the received data  $\mathbf{z} = \mathbf{s} + \mathbf{c}$ , where  $\mathbf{c}$  represents the additive unobserved complex disturbance (noise/clutter) random vector. The target signal  $\mathbf{s}$  is modeled as  $\mathbf{s} = \alpha \mathbf{p}$  where  $\mathbf{p}$  (generally called target vector response or Doppler steering vector) is the transmitted known radar pulse vector and  $\alpha = \gamma e^{j\phi} \in \mathbb{C}$  is an unknown signal parameter accounting for both channel propagation effects and the target backscattering.  $\alpha$  can be modeled as an unknown deterministic parameter or as a random variable depending on the application at hand. When modeled as a random quantity,  $\alpha$  is assumed to be a circular Gaussian random variable  $\alpha \sim \mathcal{CN}(0, \sigma_\alpha^2)$  where the amplitude  $\gamma$  is Rayleigh distributed and the phase  $\phi$  is uniformly distributed in  $[0, 2\pi]$  and independent of  $\gamma$ . Regarding the complex noise vector  $\mathbf{c}$ , it has been successfully modeled as a zero-mean CES distributed random vector with covariance matrix  $\mathbf{M} = \sigma^2 \Sigma$ , where  $\Sigma$  and  $\sigma^2$  represent the unknown scatter matrix and the unknown statistical noise power. In the following,  $\mathbf{c}$  is modeled as a complex  $t$ -distributed random vector.

The target detection problem can be expressed as a composite binary hypothesis testing problem:

$$H_0 : |\alpha| = 0 \quad \text{vs.} \quad H_1 : |\alpha| > 0, \quad (4.117)$$

or, more explicitly as:

$$\begin{cases} H_0 : \mathbf{z} = \mathbf{c}, & \mathbf{x}_m = \mathbf{c}_m, \quad m = 1, \dots, M, \\ H_1 : \mathbf{z} = \alpha \mathbf{p} + \mathbf{c}, & \mathbf{x}_m = \mathbf{c}_m, \quad m = 1, \dots, M, \end{cases} \quad (4.118)$$

where the secondary data  $\{\mathbf{x}_m\}_{m=1}^M$  are assumed to be iid and can be used to estimate the scatter matrix and share the same distribution as the clutter  $\mathbf{c}$  in the primary data vector under test.

### 4.7.1 THE ANMF DETECTOR

The NMF has been proposed, e.g., in Refs. [54,55,63–66], and can be expressed as:

$$\Lambda_{NMF} \equiv \Lambda_{NMF}(\mathbf{z}, \Sigma) = \frac{|\mathbf{p}^H \Sigma^{-1} \mathbf{z}|^2}{(\mathbf{p}^H \Sigma^{-1} \mathbf{p})(\mathbf{z}^H \Sigma^{-1} \mathbf{z})}, \quad (4.119)$$

where the scatter matrix  $\Sigma$  is assumed to be perfectly known.

An important feature of the detector in Eq. (4.119) is the invariance under scalar multiplies of  $\mathbf{x}$ . In particular, the distribution of the test statistic  $\Lambda_{NMF}$  under the hypothesis  $H_0$  is independent of the unknown average noise power  $\sigma^2$  or the functional form of the particular CES distribution of the noise, i.e., the NMF is a distribution-free detector under  $H_0$ . The proof of this property can be found in Ref. [21]. Moreover, it can be shown that  $\Lambda_{NMF}|H_0$  follows a Beta distribution:

$$\Lambda_{NMF}|H_0 \sim \text{Beta}(1, N - 1), \quad (4.120)$$

where  $\text{Beta}(x; \alpha, \beta) = (x^{\alpha-1}(1-x)^{\beta-1})/\text{B}(\alpha, \beta)$ ,  $N$  is the dimension of the data vector, and  $\text{B}(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$ .

It is clear that the NMF cannot be used in practical applications where the scatter matrix  $\Sigma$  of the data vectors is generally unknown. So, in the following simulations, we aim at investigating the detection performance of the adaptive NMF (ANMF) when  $\Sigma$  is replaced by (1)  $\hat{\Sigma}_{CMML}$ , i.e., the CMMML estimator derived in Eq. (4.95) or (2)  $\hat{\Sigma}_T$ , i.e., its min-max (over the CES distributions) robust C-Tyler estimator. Hence, we get the two adaptive detectors:

$$\Lambda_{ANMF-CMML} \equiv \Lambda_{ANMF-CMML}(\mathbf{z}, \hat{\Sigma}_{CMML}) = \frac{|\mathbf{p}^H \hat{\Sigma}_{CMML}^{-1} \mathbf{z}|^2}{(\mathbf{p}^H \hat{\Sigma}_{CMML}^{-1} \mathbf{p})(\mathbf{z}^H \hat{\Sigma}_{CMML}^{-1} \mathbf{z})} \quad (4.121)$$

$$\Lambda_{ANMF-C-Tyler} \equiv \Lambda_{ANMF-C-Tyler}(\mathbf{z}, \hat{\Sigma}_T) = \frac{|\mathbf{p}^H \hat{\Sigma}_T^{-1} \mathbf{z}|^2}{(\mathbf{p}^H \hat{\Sigma}_T^{-1} \mathbf{p})(\mathbf{z}^H \hat{\Sigma}_T^{-1} \mathbf{z})}. \quad (4.122)$$

As a consequence of the consistency of both the CMMML and Tyler's estimators, the resulting adaptive test statistic  $\Lambda_{ANMF}$  will have asymptotically (i.e., for large  $M$ ) a  $\text{Beta}(1, N - 1)$  distribution. Hence, for large  $M$ ,  $\Lambda_{ANMF}$  is (approximately) constant false alarm rate (CFAR) w.r.t.  $\Sigma$ , as desired [21]. Further discussions on the asymptotic properties of the  $\Lambda_{ANMF}$  can be found in Refs. [71, 72].

In the following, the performance of the two  $\Lambda_{ANMF}$  detectors in Eqs. (4.121) and (4.122) is compared with that of the clairvoyant *linear threshold detector* (LTD) [42], i.e., the GLRT (with respect to the unknown complex signal amplitude  $\alpha$ ) for  $t$ -distributed data under the assumption of known scatter matrix and known shape and scale parameters. In particular, the clairvoyant LTD has been derived in Ref. [42]:

$$\Lambda_{LTD} \equiv \Lambda_{LTD}(\mathbf{z}, \Sigma, \lambda, \eta) = \frac{|\mathbf{p}^H \Sigma^{-1} \mathbf{z}|^2}{(\mathbf{p}^H \Sigma^{-1} \mathbf{p})(\mathbf{z}^H \Sigma^{-1} \mathbf{z} + \lambda/\eta)}. \quad (4.123)$$

The detection performance of the  $\Lambda_{LTD}$  provides a useful upper bound to the performance of any adaptive detection algorithms, and in particular to  $\Lambda_{ANMF-CMML}$  and

$\Lambda_{ANMF-C-Tyler}$ . In our simulation, the detectors are compared in terms of (i) CFAR property w.r.t. the scatter matrix and the extra parameters, (ii) probability of detection ( $P_D$ ) as function of the signal-to-disturbance power ratio (SDR), and (iii) receiver operating characteristic (ROC) curves.

### 4.7.2 DETECTION PERFORMANCE

The detection performance of the NMF detector, which exploits either the CMML estimator or the C-Tyler estimator, is investigated by deriving by Monte Carlo simulation the following curves:

1. The probability of false alarm ( $P_{FA}$ ) as function of the one-lag coefficient  $\rho$  (Fig. 4.17). This verifies the CFAR property of the  $\Lambda_{ANMF-CMML}$  (Eq. 4.121) and the  $\Lambda_{ANMF-C-Tyler}$  (Eq. 4.122) w.r.t. the correlation shape. Simulation parameters:  $N=16$ ,  $M=3N$ ,  $\lambda=3$ ,  $\eta=1$ ,  $K=4$ . The detection thresholds have been set to achieve a nominal  $P_{FA}$  of  $10^{-3}$ . The number of Monte Carlo runs is  $10^6$ .
2. The probability of false alarm ( $P_{FA}$ ) as function of the shape parameter  $\lambda$  of the true complex  $t$ -distribution, i.e., for different spikiness levels (Fig. 4.18). This allows us to investigate the CFAR property of the two ANMFs w.r.t. the spikiness of the data. Simulation parameters:  $N=16$ ,  $M=3N$ ,  $\rho=0.8$ ,  $\eta=1$ ,  $K=4$ . The

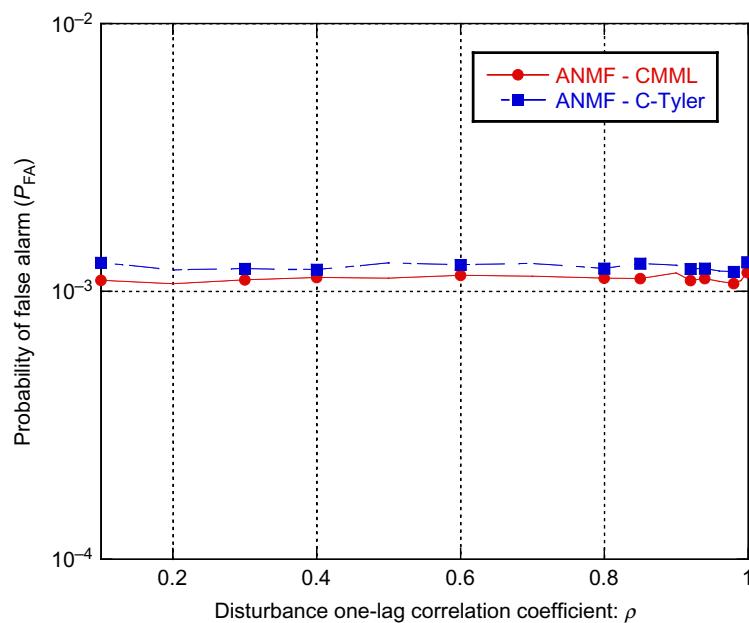
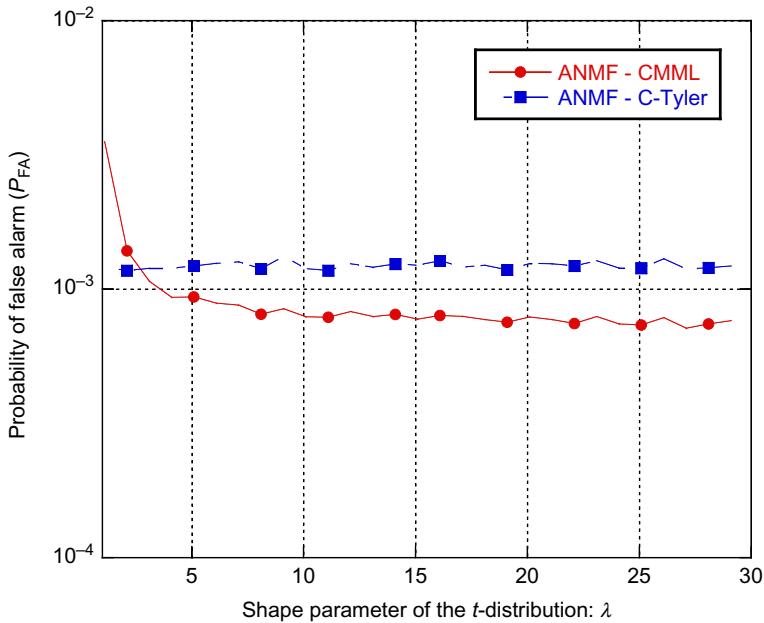


FIG. 4.17

Probability of false alarm vs. disturbance one-lag correlation coefficient  $\rho$ .

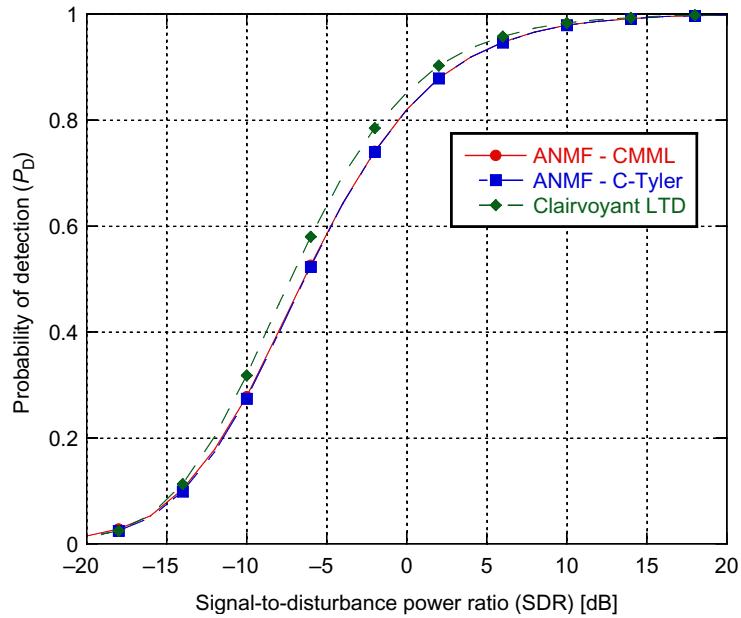
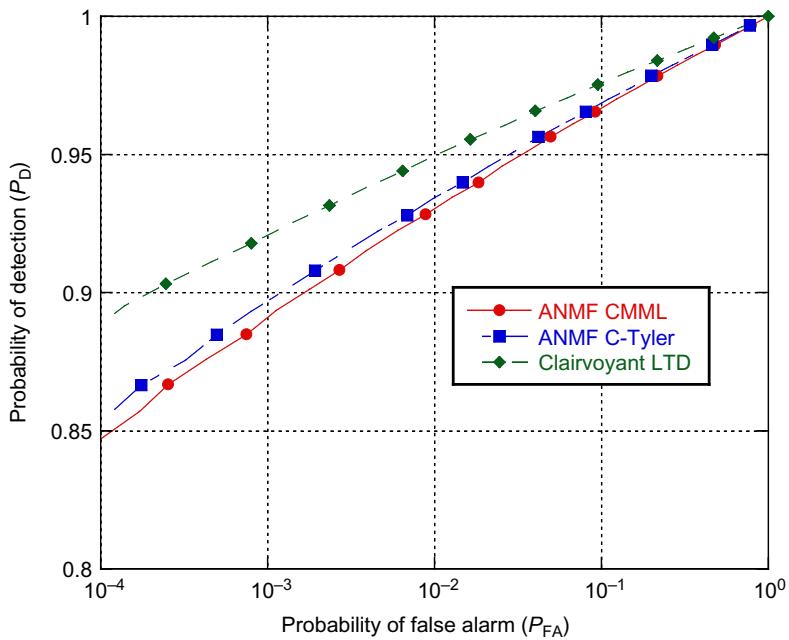
**FIG. 4.18**

Probability of false alarm vs.  $\lambda$ .

detection thresholds have been set to achieve a nominal  $P_{FA}$  of  $10^{-3}$ . The number of Monte Carlo runs is  $10^6$ .

3. The probability of detection ( $P_D$ ) as function of the SDR (Fig. 4.19). We also plot the performance of the clairvoyant LTD  $\Lambda_{LTD}$  in Eq. (4.123) that represent an upper bound to the performance achievable by  $\Lambda_{ANMF-CMMI}$  and  $\Lambda_{ANMF-C-Tyler}$ . Simulation parameters:  $N=16$ ,  $M=3N$ ,  $\rho=0.8$ ,  $\lambda=1$ ,  $\eta=1$ ,  $K=4$ . The detection thresholds have been set to achieve a nominal  $P_{FA}$  of  $10^{-3}$ . Moreover,  $\alpha \sim CN(0, \sigma_\alpha^2)$  where  $\sigma_\alpha^2$  varies according to the desired value of the SDR. The number of Monte Carlo runs is  $10^6$ .
4. The ROC curves (Fig. 4.20). The simulation parameters are  $N=16$ ,  $M=3N$ ,  $\rho=0.8$ ,  $\lambda=3$ ,  $\eta=1$ ,  $K=4$ . As before,  $\alpha \sim CN(0, \sigma_\alpha^2)$  where  $\sigma_\alpha^2$  is set to have an SDR equal to 3 dB. The number of Monte Carlo runs is  $10^6$ . Also in this case, as upper bound, the performance of  $\Lambda_{LTD}$  is reported.

As we can see from Fig. 4.17, both the ANMF detectors are (approximately) CFAR with respect to the disturbance one-lag correlation coefficient  $\rho$ . Their  $P_{FA}$  curves are almost constant and very close to the nominal  $P_{FA}$  value of  $10^{-3}$ . A similar behavior can be observed in Fig. 4.18, where the  $P_{FA}$  curves have been evaluated as function of  $\lambda$ . It can be noted that both  $\Lambda_{ANMF-CMMI}$  and  $\Lambda_{ANMF-C-Tyler}$  are CFAR detector w.r.t. the data spikiness, except for very low values  $\lambda$ , where the CMMI estimator

**FIG. 4.19**Probability of detection vs. SDR for  $P_{FA} = 10^{-3}$ .**FIG. 4.20**

Receiver operating characteristic (ROC) curves.

has large estimation losses (see Fig. 4.13) and the  $P_{\text{FA}}$  of  $\Lambda_{\text{ANMF-CMML}}$  rapidly increases. Finally, in Figs. 4.19 and 4.20, the  $P_{\text{D}}$  vs. the SDR and the ROC curves of  $\Lambda_{\text{ANMF-CMML}}$  and  $\Lambda_{\text{ANMF-C-Tyler}}$  are shown. For the sake of comparison, we also plot the detection performance of the clairvoyant GLRT for  $t$ -distributed data, i.e., the  $\Lambda_{\text{LTD}}$  of Eq. (4.123) where  $\Sigma$ ,  $\lambda$ , and  $\eta$  are assumed to be *a priori* known. The performance of  $\Lambda_{\text{ANMF-CMML}}$  and  $\Lambda_{\text{ANMF-C-Tyler}}$  is pretty close to that of the clairvoyant detector  $\Lambda_{\text{LTD}}$ . However, the adaptation losses increase when  $P_{\text{FA}}$  gets lower.

---

## 4.8 CONCLUSIONS

In practical applications, a certain amount of mismatch between the true and the assumed statistical data model is inevitable. Several authors in the statistical literature have shown how the classical tools of the estimation theory can be generalized to a mismatched scenario. In the first part of this chapter, a comprehensive review of the main contributions to the mismatched maximum likelihood theory has been proposed and discussed. A CRB under mismatched condition, i.e., the MCRB, was described and the behavior of the MML estimator was investigated. In particular, we showed that the MML estimator is asymptotically MS-unbiased and its error covariance matrix asymptotically equates the MCRB. Moreover, a constrained version of MCRB is also described. In the second part of the chapter, we showed how to apply these results to a well-known problem in radar signal processing, i.e., the problem of estimating the disturbance covariance matrix for adaptive radar target detection. We addressed this problem by putting it in the more general context of the scatter matrix estimation of CES distributed random vectors under data mismodeling. Two relevant scenarios have been considered. In the first one, the extra parameters of the particular CES distributions at hand are assumed *a priori* known. This allowed us to investigate the performance losses in the scatter matrix estimation due to a wrong specification of the functional form of the density generator. In the second scenario, we investigated the more realistic case where all the parameters are unknown and should be jointly estimated. We finished the chapter with an analysis of an adaptive detection algorithm, i.e., the ANMF which exploits either the MML estimator or the robust Tyler's estimator of the disturbance scatter matrix. The respective detection performance was compared with that of the clairvoyant GLRT detector that relies on the correct model assumption and knows *a priori* the disturbance parameters.

The mismatched approach to signal processing problem is a relatively new research field and, even though it promises huge opportunities for applicability to a plethora of different signal processing areas, many aspects still remain unresolved. In Ref. [13] for example, a generalization of the misspecified approach to the Bhattacharyya bound, to the Barankin bound, and to the Bobrovsky-Mayer-Wolf-Zakai bound has been proposed, but much work remains to be done. More importantly, in Ref. [16], Richmond posed the bases to apply the mismatched approach to the derivation of Bayesian bounds, but the path to reach a complete misspecified Bayesian estimation theory is still long.

---

## APPENDIX A A GENERALIZATION OF THE SLEPIAN FORMULA UNDER MISSPECIFICATION

In this appendix, we report the misspecified version of the Slepian formula [73] proposed by Richmond and Horowitz in their seminal paper [13]. For the sake of clarity, in the following we use the same notation as in Ref. [13]. In particular,  $(\cdot)^H$  and  $(\cdot)^*$  denote the Hermitian and the complex conjugate operators.

Let the complex data vector  $\mathbf{x} \in \mathbb{C}^N$  have a true complex Gaussian distribution such that  $\mathbf{x} \sim p_X(\mathbf{x}) = \mathcal{CN}(\mathbf{d}, \mathbf{B})$  where  $\mathbf{d}$  and  $\mathbf{B}$  denote the (possibly complex) true mean value vector and the true covariance matrix. Let the assumed distribution for the data vector be another complex Gaussian distribution such that  $\mathbf{x} \sim f_X(\mathbf{x}; \boldsymbol{\theta}) = \mathcal{CN}(\mathbf{r}(\boldsymbol{\theta}), \mathbf{R})$  where the (possibly complex) assumed mean value  $\mathbf{r}(\boldsymbol{\theta})$  is parameterized by a parameter vector  $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d$  and  $\mathbf{R}$  denotes the (possibly complex) assumed covariance matrix generally different from  $\mathbf{B}$ . For the generalization to complex parameter vector we refer the reader to Ref. [13]. It can be noted that the assumed mean value  $\mathbf{r}(\boldsymbol{\theta})$  may be different from the true one,  $\mathbf{d}$ , for every  $\boldsymbol{\theta} \in \Theta$ . Under these assumptions, the matrices  $\mathbf{A}_{\boldsymbol{\theta}}$  and  $\mathbf{B}_{\boldsymbol{\theta}}$  in Eqs. (4.1) and (4.4) can be written explicitly through the so-called misspecified Slepian formulae [13]. In particular, we have:

$$\begin{aligned} [\mathbf{A}_{\boldsymbol{\theta}}]_{i,k} = & -\frac{\partial \mathbf{r}^H(\boldsymbol{\theta})}{\partial \theta_i} \mathbf{R}^{-1} \frac{\partial \mathbf{r}(\boldsymbol{\theta})}{\partial \theta_k} - \frac{\partial \mathbf{r}^H(\boldsymbol{\theta})}{\partial \theta_k} \mathbf{R}^{-1} \frac{\partial \mathbf{r}(\boldsymbol{\theta})}{\partial \theta_i} + \\ & + \frac{\partial^2 \mathbf{r}^H(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_k} \mathbf{R}^{-1} (\mathbf{d} - \mathbf{r}(\boldsymbol{\theta})) + (\mathbf{d} - \mathbf{r}(\boldsymbol{\theta}))^H \mathbf{R}^{-1} \frac{\partial^2 \mathbf{r}(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_k}, \end{aligned} \quad (\text{A.1})$$

$$[\mathbf{B}_{\boldsymbol{\theta}}]_{i,k} = \frac{\partial \mathbf{r}^H(\boldsymbol{\theta})}{\partial \theta_i} \mathbf{R}^{-1} \mathbf{B} \mathbf{R}^{-1} \frac{\partial \mathbf{r}(\boldsymbol{\theta})}{\partial \theta_k} + \frac{\partial \mathbf{r}^T(\boldsymbol{\theta})}{\partial \theta_i} (\mathbf{R}^{-1})^* \mathbf{B}^* (\mathbf{R}^{-1})^* \frac{\partial \mathbf{r}^*(\boldsymbol{\theta})}{\partial \theta_k}. \quad (\text{A.2})$$

These expressions have been used in Ref. [13] and in Ref. [15] to evaluate the MCRB in Eqs. (4.9) and (4.18) for the DOA estimation problem with array position errors.

---

## APPENDIX B A GENERALIZATION OF THE BANGS FORMULA UNDER MISSPECIFICATION

In this appendix, we provide a misspecified version of the Bangs [74] formula derived, also in this case, by Richmond and Horowitz in Ref. [13]. As in Appendix A, let the complex data vector  $\mathbf{x} \in \mathbb{C}^N$  have a true complex Gaussian distribution such that  $\mathbf{x} \sim p_X(\mathbf{x}) = \mathcal{CN}(\mathbf{d}, \mathbf{B})$  where  $\mathbf{d}$  and  $\mathbf{B}$  denote the (possibly complex) true mean value vector and the true covariance matrix. Let the assumed distribution for the data vector be another complex Gaussian distribution such that  $\mathbf{x} \sim f_X(\mathbf{x}; \boldsymbol{\theta}) = \mathcal{CN}(\mathbf{r}, \mathbf{R}(\boldsymbol{\theta}))$  where  $\mathbf{r}$  is the (possibly complex) assumed mean value,

generally different form  $\mathbf{d}$ , and  $\mathbf{R}(\boldsymbol{\theta}) \triangleq \mathbf{R}_{\boldsymbol{\theta}}$  denotes the (possibly complex) assumed covariance matrix parameterized by a parameter vector  $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d$  and generally different from  $\mathbf{B} \forall \boldsymbol{\theta}$ . For the generalization to complex parameter vector we refer the reader to Ref. [13]. Under these assumptions, the matrices  $\mathbf{A}_{\boldsymbol{\theta}}$  and  $\mathbf{B}_{\boldsymbol{\theta}}$  in Eqs. (4.1) and (4.4) can be written explicitly through the so-called misspecified Bangs formulae [13]. In particular, we have:

$$\begin{aligned} [\mathbf{A}_{\boldsymbol{\theta}}]_{i,k} &= \text{tr} \left( \mathbf{R}_{\boldsymbol{\theta}}^{-1} \frac{\partial^2 \mathbf{R}_{\boldsymbol{\theta}}}{\partial \theta_i \partial \theta_k} \left( \mathbf{R}_{\boldsymbol{\theta}}^{-1} (\mathbf{B} + (\mathbf{d} - \mathbf{r}(\boldsymbol{\theta}))(\mathbf{d} - \mathbf{r}(\boldsymbol{\theta}))^H) - \mathbf{I}_N \right) \right) \\ &\quad - \text{tr} \left( \mathbf{R}_{\boldsymbol{\theta}}^{-1} \frac{\partial \mathbf{R}_{\boldsymbol{\theta}}}{\partial \theta_k} \mathbf{R}_{\boldsymbol{\theta}}^{-1} \frac{\partial \mathbf{R}_{\boldsymbol{\theta}}}{\partial \theta_i} \left( \mathbf{R}_{\boldsymbol{\theta}}^{-1} (\mathbf{B} + (\mathbf{d} - \mathbf{r}(\boldsymbol{\theta}))(\mathbf{d} - \mathbf{r}(\boldsymbol{\theta}))^H) - \mathbf{I}_N \right) \right) \\ &\quad - \text{tr} \left( \mathbf{R}_{\boldsymbol{\theta}}^{-1} \frac{\partial \mathbf{R}_{\boldsymbol{\theta}}}{\partial \theta_i} \mathbf{R}_{\boldsymbol{\theta}}^{-1} \frac{\partial \mathbf{R}_{\boldsymbol{\theta}}}{\partial \theta_k} \mathbf{R}_{\boldsymbol{\theta}}^{-1} (\mathbf{B} + (\mathbf{d} - \mathbf{r}(\boldsymbol{\theta}))(\mathbf{d} - \mathbf{r}(\boldsymbol{\theta}))^H) \right), \end{aligned} \quad (\text{B.1})$$

$$\begin{aligned} [\mathbf{B}_{\boldsymbol{\theta}}]_{i,k} &= \text{tr} \left( \mathbf{R}_{\boldsymbol{\theta}}^{-1} \frac{\partial \mathbf{R}_{\boldsymbol{\theta}}}{\partial \theta_i} \mathbf{R}_{\boldsymbol{\theta}}^{-1} \mathbf{B} \mathbf{R}_{\boldsymbol{\theta}}^{-1} \frac{\partial \mathbf{R}_{\boldsymbol{\theta}}}{\partial \theta_k} \mathbf{R}_{\boldsymbol{\theta}}^{-1} (\mathbf{B} + (\mathbf{d} - \mathbf{r}(\boldsymbol{\theta}))(\mathbf{d} - \mathbf{r}(\boldsymbol{\theta}))^H) \right) \\ &\quad + (\mathbf{d} - \mathbf{r}(\boldsymbol{\theta}))^H \mathbf{R}_{\boldsymbol{\theta}}^{-1} \frac{\partial \mathbf{R}_{\boldsymbol{\theta}}}{\partial \theta_k} \mathbf{R}_{\boldsymbol{\theta}}^{-1} \mathbf{B} \mathbf{R}_{\boldsymbol{\theta}}^{-1} \frac{\partial \mathbf{R}_{\boldsymbol{\theta}}}{\partial \theta_i} \mathbf{R}_{\boldsymbol{\theta}}^{-1} (\mathbf{d} - \mathbf{r}(\boldsymbol{\theta})), \end{aligned} \quad (\text{B.2})$$

where  $\mathbf{I}_N$  is the identity matrix of order  $N$ .

## APPENDIX C COMPACT EXPRESSION FOR THE MCRB IN THE CES FAMILY

In this appendix, we derive a compact expression useful to evaluate the MCRB for the scatter matrix estimation in the family of CES distribution. This expression follows directly from the results obtained in Ref. [46]. We assume that both the true distribution  $p_X(\mathbf{x})$  (that implicitly depends on the true scatter matrix  $\bar{\Sigma}$ , then according to the notation used before,  $\bar{\boldsymbol{\theta}} = \text{vecs}(\bar{\Sigma})$ ) and the assumed distribution  $f_X(\mathbf{x}; \boldsymbol{\Sigma})$  belong to the zero-mean CES distribution class, as shown in Eqs. (4.48) and (4.49). Moreover, we define  $Q \triangleq \mathbf{x}^H \boldsymbol{\Sigma}^{-1} \mathbf{x}$  as in Eq. (4.54).

### COMPACT EXPRESSION FOR THE MATRIX $\mathbf{B}_{\bar{\boldsymbol{\theta}}}$

In Ref. [46] the matrix  $\mathbf{B}_{\bar{\boldsymbol{\theta}}}$  has been obtained element by element as:

$$\begin{aligned} [\mathbf{B}_{\bar{\boldsymbol{\theta}}}]_{ij} &= E_p \left\{ \frac{\partial \ln f_X(\mathbf{x}; \boldsymbol{\theta})}{\partial \theta_i} \frac{\partial \ln f_X(\mathbf{x}; \boldsymbol{\theta})}{\partial \theta_j} \right\} \\ &= \left( 1 + \frac{2}{N} E \left\{ Q \frac{\partial \ln g(Q)}{\partial Q} \right\} + \frac{1}{N(N+1)} E \left\{ Q^2 \left( \frac{\partial \ln g(Q)}{\partial Q} \right)^2 \right\} \right) \text{tr}(\bar{\boldsymbol{\Sigma}}^{-1} \mathbf{A}_i) \text{tr}(\bar{\boldsymbol{\Sigma}}^{-1} \mathbf{A}_j) \\ &\quad + \frac{1}{N(N+1)} E \left\{ Q^2 \left( \frac{\partial \ln g(Q)}{\partial Q} \right)^2 \right\} \text{tr}(\bar{\boldsymbol{\Sigma}}^{-1} \mathbf{A}_i \bar{\boldsymbol{\Sigma}}^{-1} \mathbf{A}_j), \end{aligned} \quad (\text{C.1})$$

where  $\mathbf{A}_i = \partial\boldsymbol{\Sigma}/\partial\theta_i$  is a symmetric 0-1 matrix.

For notation simplicity, we define:

$$B_1 = 1 + \frac{2}{N} E \left\{ Q \frac{\partial \ln g(Q)}{\partial Q} \right\} + \frac{1}{N(N+1)} E \left\{ Q^2 \left( \frac{\partial \ln g(Q)}{\partial Q} \right)^2 \right\}, \quad (\text{C.2})$$

$$B_2 = \frac{1}{N(N+1)} E \left\{ Q^2 \left( \frac{\partial \ln g(Q)}{\partial Q} \right)^2 \right\}. \quad (\text{C.3})$$

By using the properties of the vec operator, of the Duplication matrix  $\mathbf{D}_N$  and of the Kronecker product [47,48], we have:

$$\mathbf{B}_{\bar{\theta}} = \mathbf{D}_N^T \left[ B_1 \text{vec}(\boldsymbol{\Sigma}^{-1}) \text{vec}(\boldsymbol{\Sigma}^{-1})^T + B_2 \boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1} \right] \mathbf{D}_N \quad (\text{C.4})$$

### COMPACT EXPRESSION FOR THE MATRIX $\mathbf{A}_{\bar{\theta}}$

In Ref. [46] the matrix  $\mathbf{A}_{\bar{\theta}}$  has been obtained element by element as:

$$\begin{aligned} [\mathbf{A}_{\bar{\theta}}]_{ij} &= E_p \left\{ \frac{\partial^2 \ln f_X(\mathbf{x}; \boldsymbol{\theta})}{\partial \theta_i \partial \theta_j} \right\} \\ &= \left( 1 + \frac{2}{N} E \left\{ Q \frac{\partial \ln g(Q)}{\partial Q} \right\} + \frac{1}{N(N+1)} E \left\{ Q^2 \frac{\partial^2 \ln g(Q)}{\partial Q^2} \right\} \right) \text{tr}(\boldsymbol{\Sigma}^{-1} \mathbf{A}_i \boldsymbol{\Sigma}^{-1} \mathbf{A}_j) + \quad (\text{C.5}) \\ &\quad + \frac{1}{N(N+1)} E \left\{ Q^2 \frac{\partial^2 \ln g(Q)}{\partial Q^2} \right\} \text{tr}(\boldsymbol{\Sigma}^{-1} \mathbf{A}_i) \text{tr}(\boldsymbol{\Sigma}^{-1} \mathbf{A}_j) \end{aligned}$$

For notation simplicity, we define:

$$A_2 = 1 + \frac{2}{N} E \left\{ Q \frac{\partial \ln g(Q)}{\partial Q} \right\} + \frac{1}{N(N+1)} E \left\{ Q^2 \frac{\partial^2 \ln g(Q)}{\partial Q^2} \right\}, \quad (\text{C.6})$$

$$A_1 = \frac{1}{N(N+1)} E \left\{ Q^2 \frac{\partial^2 \ln g(Q)}{\partial Q^2} \right\}. \quad (\text{C.7})$$

Finally, as for the matrix  $\mathbf{B}_{\bar{\theta}}$ , we have:

$$\mathbf{A}_{\bar{\theta}} = \mathbf{D}_N^T \left[ A_1 \text{vec}(\boldsymbol{\Sigma}^{-1}) \text{vec}(\boldsymbol{\Sigma}^{-1})^T + A_2 \boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1} \right] \mathbf{D}_N. \quad (\text{C.8})$$

By using the Sherman-Morrison formula, we can express the inverse of the matrix  $\mathbf{A}$  as follows:

$$\mathbf{A}_{\bar{\theta}}^{-1} = \mathbf{D}_N^\dagger \left[ \frac{1}{A_2} \boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma} - \frac{A_1}{A_2(A_2 + NA_1)} \text{vec}(\boldsymbol{\Sigma}) \text{vec}(\boldsymbol{\Sigma})^T \right] \left( \mathbf{D}_N^\dagger \right)^T. \quad (\text{C.9})$$

## COMPACT EXPRESSION FOR THE MCRB, $\text{MCRB}(\bar{\theta}) = \mathbf{M}^{-1} \mathbf{A}_{\bar{\theta}}^{-1} \mathbf{B}_{\bar{\theta}} \mathbf{A}_{\bar{\theta}}^{-1}$ (WITH $\mathbf{R} = \mathbf{0}$ )

$$\begin{aligned}\text{MCRB}(\bar{\theta}) &= \frac{1}{M} \mathbf{A}_{\bar{\theta}}^{-1} \mathbf{B}_{\bar{\theta}} \mathbf{A}_{\bar{\theta}}^{-1} \\ &= \frac{1}{M} \mathbf{D}_N^\dagger \left[ \frac{B_2}{A_2^2} \bar{\Sigma} \otimes \bar{\Sigma} + \left( \frac{B_1}{A_2^2} - \frac{2A_1(B_2 + NB_1)}{A_2(A_2 + NA_1)} + \frac{2NA_1^2(B_2 + NB_1)}{A_2^2(A_2 + NA_1)^2} \right) \text{vec}(\bar{\Sigma}) \text{vec}(\bar{\Sigma})^T \right] (\mathbf{D}_N^\dagger)^T.\end{aligned}\quad (\text{C.10})$$

## REFERENCES

- [1] P.J. Huber, The behavior of maximum likelihood estimates under nonstandard conditions, in: Proc. of the Fifth Berkeley Symposium in Mathematical Statistics and Probability, University of California Press, Berkley, 1967.
- [2] H. White, Maximum likelihood estimation of misspecified models, *Econometrica* 50 (1982) 1–25.
- [3] H. White, Consequences and detection of misspecified nonlinear regression models, *J. Am. Stat. Assoc.* 76 (1981) 419–433.
- [4] H. White, *Estimation, Inference and Specification Analysis*, Econometric Society Monographs, Cambridge University Press, Cambridge, United Kingdom, 1996.
- [5] Q.H. Vuong, Cramér-Rao bounds for misspecified models, Working Paper 652, Division of the Humanities and Social Sciences, Caltech, 1986. Available at: <https://www.hss.caltech.edu/content/cramer-rao-bounds-misspecified-models>.
- [6] T.B. Fomby, R.C. Hill, *Maximum-Likelihood Estimation of Misspecified Models: Twenty Years Later*, Elsevier Ltd, Kidington, Oxford, UK, 2003.
- [7] Y. Noam, J. Tabrikian, Marginal likelihood for estimation and detection theory, *IEEE Trans. Signal Process.* 55 (8) (2007) 3963–3974.
- [8] W. Xu, A.B. Bagherer, K.L. Bell, A bound on mean-square estimation error with background parameter mismatch, *IEEE Trans. Inf. Theory* 50 (4) (2004). 621, 632.
- [9] A. Gusi-Amigó, P. Closas, A. Mallat, L. Vandendorpe, Ziv-Zakai lower bound for UWB based TOA estimation with unknown interference, in: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Florence, 2014, pp. 6504–6508.
- [10] A. Gusi-Amigó, P. Closas, L. Vandendorpe, Mean square error performance of sample mean and sample median estimators, in: *IEEE Statistical Signal Processing Workshop (SSP)*, Palma de Mallorca, 2016, pp. 1–5.
- [11] A. Gusi-Amigó, P. Closas, L. Vandendorpe, Mean Square Error bounds for parameter estimation under model misspecification, 2015, arXiv:1511.03982 [math.ST].
- [12] C.D. Richmond, L.L. Horowitz, Parameter bounds under misspecified models, in: *Conference on Signals, Systems and Computers*, 3–6 November 2013, Asilomar, 2013, pp. 176–180.
- [13] C.D. Richmond, L.L. Horowitz, Parameter bounds on estimation accuracy under model misspecification, *IEEE Trans. Signal Process.* 63 (9) (2015) 2263–2278.
- [14] P.A. Parker, C.D. Richmond, Methods and bounds for waveform parameter estimation with a misspecified model, in: *49th Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, November, 2015, pp. 1702–1706.

- [15] C. Ren, M.N. El Korsos, J. Galy, E. Chaumette, P. Larzabal, A. Renaux, Performances bounds under misspecification model for MIMO radar application, in: Proc. of Eur. Signal Process. Conf. (EUSIPCO), Nice, France, 2015, pp. 514–518.
- [16] C.D. Richmond, P. Basu, Bayesian framework and radar: on misspecified bounds and radar-communication cooperation, in: IEEE Workshop on Statistical Signal Processing 2016 (SSP), Palma de Mallorca, Spain, 26–29 June, 2016.
- [17] J.M. Kantor, C.D. Richmond, B. Correll, D.W. Bliss, Prior mismatch in Bayesian direction of arrival for sparse arrays, in: IEEE Radar Conference, Philadelphia, PA, May, 2015, pp. 0811–0816.
- [18] C. Fritzsche, U. Orguner, E. Ozkan, F. Gustafsson, On the Cramér-Rao lower bound under model mismatch, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 3986–3990.
- [19] S. Fortunati, M.S. Greco, F. Gini, A lower bound for the mismatched maximum likelihood estimator, in: IEEE Radar Conf. 2015, Arlington, USA, May 11–15, 2015.
- [20] S. Fortunati, F. Gini, M.S. Greco, The misspecified Cramér-Rao bound and its application to the scatter matrix estimation in complex elliptically symmetric distributions, *IEEE Trans. Signal Process.* 64 (9) (2016) 2387–2399.
- [21] E. Ollila, D.E. Tyler, V. Koivunen, V.H. Poor, Complex elliptically symmetric distributions: survey, new results and applications, *IEEE Trans. Signal Process.* 60 (11) (2012) 5597–5625.
- [22] T.J. Rothenberg, Identification in parametric models, *Econometrica* 39 (3) (1971) 577–591.
- [23] R. Bowden, The theory of parametric identification, *Econometrica* 41 (6) (1973) 1069–1074.
- [24] S. Fortunati, F. Gini, M.S. Greco, A. Farina, A. Graziano, S. Giompapa, On the identifiability problem in the presence of random nuisance parameters, *Signal Process.* 92 (2012) 2545–2551.
- [25] P. Stoica, T. Söderström, On non singular information matrices and local identifiability, *Int. J. Control.* 36 (1982) 323–329.
- [26] P.J. Schreier, L.L. Scharf, *Statistical Signal Processing of Complex-Valued Data*, Cambridge University Press, Cambridge, United Kingdom, 2010.
- [27] A.K. Jagannatham, B.D. Rao, Cramér-Rao lower bound for constrained complex parameters, *IEEE Signal Process. Lett.* 11 (11) (2004) 875–878.
- [28] T. Menni, E. Chaumette, P. Larzabal, J.P. Barbot, New results on deterministic Cramér–Rao bounds for real and complex parameters, *IEEE Trans. Signal Process.* 60 (3) (2012) 1032–1049.
- [29] L.T. McWhorter, L.L. Scharf, Properties of quadratic covariance bounds, in: Conference Record of the Twenty-Seventh Asilomar Conference on Signals, Systems and Computers, vol. 2, 1–3 November 1993, 1993, pp. 1176–1180.
- [30] J.D. Gorman, A.O. Hero, Lower bounds for parametric estimation with constraints, *IEEE Trans. Inf. Theory* 6 (6) (1990) 1285–1301.
- [31] T.L. Marzetta, A simple derivation of the constrained multiple parameter Cramér-Rao bound, *IEEE Trans. Signal Process.* 41 (1993) 2247–2249.
- [32] P. Stoica, B.C. Ng, On the Cramér-Rao bound under parametric constraints, *IEEE Signal Process. Lett.* 5 (7) (1998) 177–179.
- [33] T.J. Moore, R.J. Kozick, B.M. Sadler, The constrained Cramér–Rao bound from the perspective of fitting a model, *IEEE Signal Process. Lett.* 14 (8) (2007) 564–567.
- [34] S. Fortunati, F. Gini, M.S. Greco, The constrained misspecified Cramér-Rao bound, *IEEE Signal Process. Lett.* 23 (5) (2016) 718–721.
- [35] M. Spivak, *Calculus on Manifolds*, Addison-Wesley, Reading, MA, 1965.

- [36] T.M. Cover, J.A. Thomas, *Elements of Information Theory*, John Wiley and Sons, New York, NY, 1991.
- [37] C.D. Richmond, PDF's, confidence regions, and relevant statistics for a class of sample covariance-based array processors, *IEEE Trans. Signal Process.* 44 (7) (1996) 1779–1793.
- [38] A. Balleri, A. Nehorai, J. Wang, Maximum likelihood estimation for compound-gaussian clutter with inverse gamma texture, *IEEE Trans. Aerosp. Electron. Syst.* 43 (2) (2007) 775–779.
- [39] J. Wang, A. Dogandzic, A. Nehorai, Maximum likelihood estimation of compound-gaussian clutter and target parameters, *IEEE Trans. Signal Process.* 54 (10) (2006) 3884–3898.
- [40] M.S. Greco, F. Gini, Cramér-Rao lower bounds on covariance matrix estimation for complex elliptically symmetric distributions, *IEEE Trans. Signal Process.* 61 (24) (2013) 6401–6409.
- [41] A. Younsi, M. Greco, F. Gini, A. Zoubir, Performance of the adaptive generalised matched subspace constant false alarm rate detector in non-Gaussian noise: an experimental analysis, *IET Radar Sonar Navig.* 3 (3) (2009) 195–202.
- [42] K.J. Sangston, F. Gini, M. Greco, Coherent radar detection in heavy-tailed compound-Gaussian clutter, *IEEE Trans. Aerosp. Electron. Syst.* 42 (1) (2012) 64–77.
- [43] J.R. Magnus, H. Neudecker, The commutation matrix: some properties and applications, *Ann. Stat.* 7 (1979) 381–394.
- [44] F. Gini, J.H. Michels, Performance analysis of two covariance matrix estimators in compound-gaussian clutter, *IEE Proc. Part-F* 146 (3) (1999) 133–140.
- [45] J.R. Magnus, H. Neudecker, Matrix differential calculus with applications to simple, hadamard, and kronecker products, *J. Math. Psychol.* 29 (1985) 414–492.
- [46] M.S. Greco, S. Fortunati, F. Gini, Maximum likelihood covariance matrix estimation for complex elliptically symmetric distributions under mismatched conditions, *Signal Process.* 104 (2014) 381–386.
- [47] J.R. Magnus, H. Neudecker, The elimination matrix: some lemmas and applications, *SIAM J. Algebraic Discr. Methods* 1 (1980) 422–499.
- [48] J.R. Magnus, H. Neudecker, Symmetry, 0-1 matrices and Jacobians: a review, *Economet. Theor.* 2 (1986) 157–190.
- [49] K.B. Petersen, M.S. Pedersen, The Matrix Cookbook, 2012, <http://matrixcookbook.com>.
- [50] F. Gini, M.S. Greco, Covariance matrix estimation for CFAR detection in correlated heavy tailed clutter, *Signal Process.* 82 (12) (2002) 1847–1859.
- [51] R.A. Maronna, Robust M-estimators of multivariate location and scatter, *Ann. Stat.* 4 (1) (1976) 51–67.
- [52] F. Pascal, Y. Chitour, J. Ovarlez, P. Forster, P. Larzabal, Covariance structure maximum-likelihood estimates in compound gaussian noise: existence and algorithm analysis, *IEEE Trans. Signal Process.* 56 (1) (2008). 34, 48.
- [53] D. Tyler, A distribution-free  $M$ -estimator of multivariate scatter, *Ann. Stat.* 15 (1) (1987) 234–251.
- [54] F. Gini, M. Greco, A. Farina, Clairvoyant and adaptive signal detection in non-gaussian clutter: a data-dependent threshold interpretation, *IEEE Trans. Signal Process.* 47 (6) (1999) 1522–1531.
- [55] F. Gini, M. Greco, A suboptimum approach to adaptive coherent radar detection in compound-gaussian clutter, *IEEE Trans. Aerosp. Electron. Syst.* 35 (3) (1999) 1095–1104.
- [56] F. Pascal, L. Bombrun, J.-Y. Tourneret, Y. Berthoumieu, Parameter estimation for multivariate generalized Gaussian distributions, *IEEE Trans. Signal Process.* 61 (23) (2013) 5960–5971.

- [57] I.S. Gradshteyn, I.M. Ryzhik, *Table of Integrals, Series, and Products*, seventh ed., Academic Press, Cambridge, Massachusetts, USA, 2007.
- [58] S. Fortunati, F. Gini, M.S. Greco, On scatter matrix estimation in the presence of unknown extra parameters: mismatched scenario, in: EUSIPCO 2016, Budapest, Hungary, 29 August–2 September, 2016.
- [59] C.C. Heyde, R. Morton, On constrained quasi-likelihood estimation, *Biometrika* 80 (4) (1993) 755–761.
- [60] F. Gini, Performance analysis of two structured covariance matrix estimators in compound-Gaussian clutter, *Signal Process.* 80 (2) (2000) 365–371.
- [61] S. Fortunati, F. Gini, M.S. Greco, Matched, mismatched and robust scatter matrix estimation and hypothesis testing in complex  $t$ -distributed data, *EURASIP J. Adv. Signal Process.* 2016 (2016) 123.
- [62] S. Fortunati, M.S. Greco, F. Gini, The impact of unknown extra parameters on scatter matrix estimation and detection performance in complex  $t$ -distributed data, in: IEEE Workshop on Statistical Signal Processing 2016 (SSP), Palma de Mallorca, Spain, 26–29 June, 2016.
- [63] L.L. Scharf, B. Friedlander, Matched subspace detectors, *IEEE Trans. Signal Process.* 42 (8) (1994) 2146–2157.
- [64] F. Gini, A cumulant-based adaptive technique for coherent radar detection in a mixture of K-distributed clutter and Gaussian disturbance, *IEEE Trans. Signal Process.* 45 (6) (1997) 1507–1519.
- [65] E. Conte, A. De Maio, G. Ricci, Recursive estimation of the covariance matrix of a compound-Gaussian process and its application to adaptive CFAR detection, *IEEE Trans. Signal Process.* 50 (8) (2002) 1908–1915.
- [66] F. Gini, Sub-optimum coherent radar detection in a mixture of K-distributed and Gaussian clutter, *IEE Proc. Part-F* 144 (1) (1997) 39–48.
- [67] E. Conte, A. De Maio, Mitigation techniques for non-Gaussian sea clutter, *IEEE J. Ocean. Eng.* 29 (2) (2004) 284–302.
- [68] E. Conte, A. De Maio, G. Ricci, Adaptive CFAR detection in compound-Gaussian clutter with circulant covariance matrix, *IEEE Signal Process. Lett.* 7 (3) (2000) 63–65.
- [69] E. Conte, A. De Maio, G. Ricci, Covariance matrix estimation for adaptive CFAR detection in compound-Gaussian clutter, *IEEE Trans. Aerosp. Electron. Syst.* 38 (2) (2002) 415–426.
- [70] E. Conte, A. De Maio, Exploiting persymmetry for CFAR detection in compound-Gaussian clutter, *IEEE Trans. Aerosp. Electron. Syst.* 39 (2) (2003) 719–724.
- [71] F. Pascal, J.P. Ovarlez, Asymptotic detection performance of the robust ANMF, in: 23rd European Signal Processing Conference (EUSIPCO), 2015, Nice, 2015, pp. 524–528.
- [72] J.P. Ovarlez, F. Pascal, A. Breloy, Asymptotic detection performance analysis of the robust Adaptive Normalized Matched Filter, in: IEEE CAMSAP 2015, Cancun, 2015, pp. 137–140.
- [73] D. Slepian, Estimation of signal parameters in the presence of noise, *Trans. IRE Prof. Group Inf. Theory* IT-3 (1954) 68–89.
- [74] W.J. Bangs, Array processing with generalized beamforming (PhD dissertation), Yale Univ., New Haven, CT, 1971.
- [75] A. Mennad, S. Fortunati, M.N. El Korso, A. Younsi, A.M. Zoubir, A. Renaux, Slepian-Bangs-type formulas and the related Misspecified Cramér-Rao Bounds for Complex Elliptically Symmetric Distributions, *Signal Process.* 142C (2018) 320–329.
- [76] S. Fortunati, Misspecified Cramér-Rao Bounds for Complex Unconstrained and Constrained Parameters, in: EUSIPCO 2017, Kos, Greece, 2017.

# Multistatic radar systems

# 5

**Daniel W. O'Hagan\*, Shaun R. Doughty<sup>†‡</sup>, Michael R. Inggs\***

*Radar Remote Sensing Group (RRSG), Dept. of Electrical Engineering, University of Cape Town (UCT), Cape Town, South Africa\* Dept. of Electrical and Electronic Engineering, University College London (UCL), Torrington Place, London, United Kingdom<sup>†</sup> Current affiliation: Maxeler Technologies, London, United Kingdom<sup>‡</sup>*

---

## 5.1 INTRODUCTION

The IEEE Standard Radar Definitions [1] define *multistatic radar* as: *A radar system having two or more transmitting or receiving antennas with all antennas separated by large distances when compared to the antenna sizes.*

The definitions of multistatic radar are malleable. Some authors describe it as a collection of multiple spatially diverse monostatic radars with overlapping coverage regions, called netted radar [2]. Others describe multistatic radar as collections of transmit-receive pairs [3, 4]. Hanle in [5] covers bistatic and multistatic radar definitions; however, some of these have metamorphosed over the years. Chernyak in [6] has highlighted the similarity between multistatic/multisite radar and *statistical MIMO radar*. MIMO radar, especially statistical MIMO radar, is itself contentious and this chapter will not make any particular connection between multistatic radar and MIMO radar [6]. Rather, when describing multistatic radar and techniques, stand-alone context will be provided.

This chapter will summarize the salient characteristics of multistatic radar and will cover some relevant theory.

---

## 5.2 CHARACTERISTICS OF MULTISTATIC RADAR

The spatial diversity offered by multistatic radar has potential advantages over conventional monostatic radar due to an increase in information on the target(s). Multistatic radar permits observation of targets from multiple different transmitter-receiver pairs. These distinct pairs can view different aspects of a target, thereby adding potentially useful information that could, for example, enhance target recognition.

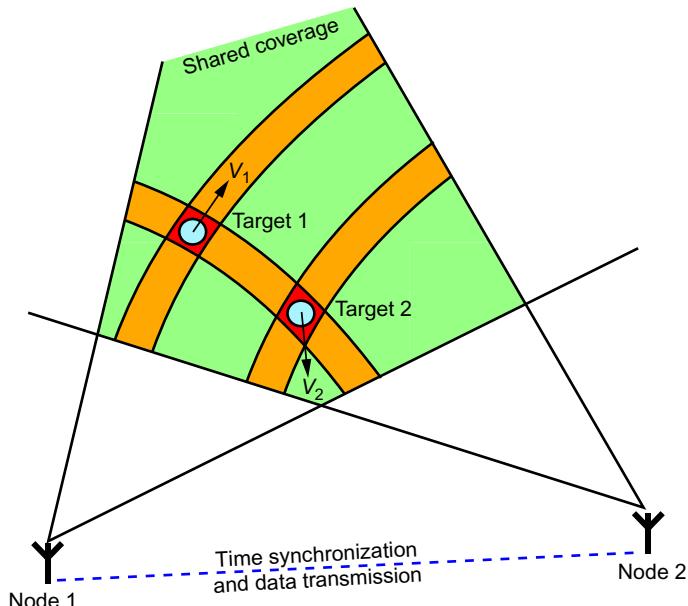
A common time reference available to distributed nodes of the multistatic radar is essential for coherent processing and information fusion. The exact precision of the timing reference can have an application dependency. Accurate time and frequency distribution throughout a widely separated network are an area of active research [7–11]. Both control and data collection from the spatially diverse sites require a suitable means of communication.

Multistatic resolution, accuracy, and detection performance are more complicated to describe than for the corresponding monostatic and bistatic parameters. These depend on the method of data fusion involved in detection, which is covered in Chapter 3 of [12]. However, an overview of the advantages of using multistatic radar is provided here.

In the case of a multistatic radar with overlapping coverage between nodes, the probability of detection for a given target could be increased as there is greater likelihood for the target to be physically closer to a transmit-receive pair.

The spatial diversity of the multistatic configuration could potentially reduce shadowing effects, which can obscure large angular sections of coverage. Spatial diversity adds value in urban, and similarly enclosed, environments where a typical radar might have a restricted field of view. It is expected that multistatic radar will have an important role to play in the emerging field of drone detection.

Multistatic radar can also be used to improve resolution and parameter estimation. Generally monostatic and bistatic radars have poorer cross-range resolution compared with down-range profiles. Fig. 5.1 shows that how the intersection of the down-range profiles within an area illuminated by multiple beamwidths (and



**FIG. 5.1**

Intersection of down-range profiles.

hence unresolvable by angle) permits Target 1 to be distinguished from Target 2—compared with Node 1 alone, where both targets are in the same resolution cell, and therefore unresolvable. Locating an object by accurately computing the time of arrival (ToA) is often referred to as multilateration. The technique typically involves at least three spatially separate ToA measurements for a ground-based system to obtain a three-dimensional target position. Multilateration using TDOA to locate targets is often used in secondary surveillance radar (SSR); however, the main added complexity in multistatic radar is the noncooperative nature of targets. That is, unlike cooperative SSR, a received radar echo does not intrinsically contain information identifying from which target it originated in a multitarget scenario.

The noise-limited accuracy of parameter estimation for monostatic and bistatic radars is proportional to both resolution and SNR. The coverage and resolution improvements described for multistatic systems may yield better estimates of target position and velocity. The spatial diversity of the system will allow the radar to see a target from several different aspects. The target velocity will contribute Doppler shifted echo signals along several different unit vectors. Therefore, the inability of monostatic radar to discern target radial velocity might be avoided in the multistatic case. With an adequate number of spatially diverse transmitter-receiver pairs, the target velocity may be fully specified as a two- or three-dimensional vector.

Multiple velocity vectors are one example of increased information that can emerge from having multistatic aspects. Additional information such as target radar cross section (RCS) variation with aspect and micro-Doppler profiles are multistatic radar measurements that can enhance target identification schemes.

The appearance of clutter will also be altered through the spatial diversity of bistatic and multistatic radars. The literature on bistatic and multistatic clutter is sparse; however, it is expected to grow in coming years, especially with the development of NeXtRAD. Sea clutter is taken as an example to expatiate on the characteristics of clutter as a function of spatial diversity. When a monostatic radar is orientated to directly face approaching sea waves, a large clutter reflectivity can be observed in the nonzero Doppler region. On the other hand, using a multistatic radar with different transmitter-receiver pairs can, in some instances, but not all, reduce the “spikiness” of the clutter and thus has value for improving target detection [12, 13].

Finally, in conflict situations a sensor should be as radio frequency (RF) silent as possible to reduce the risk of it being targeted. With passive coherent location systems, the dislocated receiver site is RF silent. However, all transmitters, whether illuminators of opportunity or otherwise, are vulnerable to antiradiation homing and physical attack regardless of being colocated with a receiver or not. In wartime it is prudent to assume that transmitter infrastructure would be targeted [14]. Multistatic radars, however, offer enhanced survivability and “graceful degradation” because of their decentralized, distributed, nature. A fault in, or destruction of, either a monostatic or bistatic transmitter or receiver would lead to a complete loss in radar functionality. From a tactical perspective, a single large transmitter would be easier to locate and destroy than several spatially distributed transmitters. In addition, it is difficult to successfully execute jamming on multiple receivers compared to a single site—especially if the receiver locations are unknown [15].

### 5.3 MULTISTATIC RADAR TECHNOLOGY ENABLERS

Advances in digital signal processing, field programmable gate array (FPGA) technology, and software defined radio (SDR) have been enablers for bistatic and multistatic radar. These developments have increasingly resulted in “traditionally” analog signal processing tasks being handled digitally, with filtering, synchronization of time and phase, deterministic triggering, and other core components of the radar design addressed by a multitude of discrete dedicated hardware platforms. Meeting the needs of evolving radar systems, this hardware logic can now be implemented programmatically on a single FPGA as multiple independent IP cores that can operate in parallel [16, 17]. FPGAs provide the ability to directly interface and control connected hardware and allow for a modular upgradable architecture whereby its internal logic can be reconfigured for task-specific applications. Through integration with technologies such as GPS, the digital back-end is able to calibrate for geolocation to achieve pointing accuracy and synchronize clocks for each node in a multistatic radar network [18].

SDR has been available for more than two decades, but the requirements of radar have only recently been met. SDR transmitter and receiver devices are now available as commercial off-the-shelf (COTS) platforms supporting quadrature signals at 16-bit resolution, sampling rates up to 250 MS/s and frequency coverage from 300 kHz to 6 GHz. Some SDRs enable portable low power radar receivers, with an antenna being the only additional requirement. Leveraging advancements in enabling technologies permits the establishment of phase coherent multistatic radar network systems.

---

### 5.4 SIGNAL PROCESSING IN MULTISTATIC RADAR

One of the key aspects of a Multistatic Radar design is the location of the decision-making process within the system. A system may aim to detect, resolve, locate, identify, and track targets of interest—each of these processes within a multistatic radar may be:

- Centralized, where maximum information is transmitted to a central location, whereupon a joint decision is made.
- Decentralized, where more decision making is first performed at each node, reducing the information transmitted further to the central location prior to a joint decision. The decentralized process is also referred to as “distributed.”

An example of a highly centralized system would be that where RF signals are directly combined to form a joint detection (such as in a phased array). An example of a highly decentralized system might be the combining of target tracks from multiple radars.

It is difficult to provide a strict classification for the degree of centralization; and perhaps unhelpful given the diversity of potential applications. Certainly hybrid

approaches (e.g., a decentralized system with feedback from a joint decision) are also possible and could provide performance benefits in certain applications.

While decentralization systems may be performance limited in some aspects (due to loss of information), they will often convey more practical advantages. These include a reduction in communications bandwidth and processing requirements at the point of joint decision making.

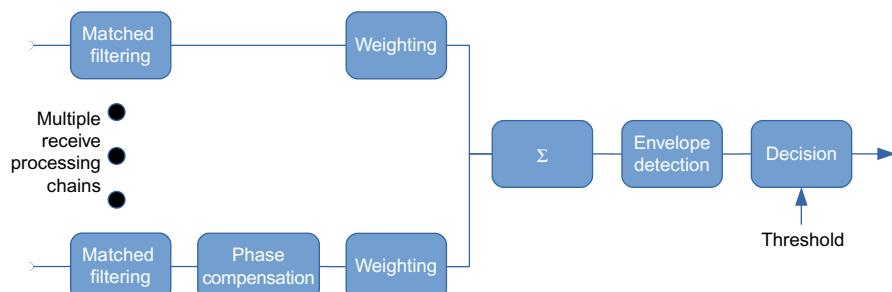
Optimizing signal processing for one system function in isolation during design will not necessarily produce a practical system. For example, Chernyak [6] derived detectors based on the generalized likelihood ratio for a signal arriving at all receive stations, in the presence of white Gaussian noise. However, direct application of these detectors gives no guarantee of resolution between multiple targets.

To characterize target resolution, attempts to use traditional tools such as ambiguity functions have been undertaken (e.g., [19]) but the large number of additional parameters in a multistatic system (e.g., the system topology) means that the ambiguity function loses much of the generalization/simplicity that makes it such a useful tool in monostatic system analysis.

## 5.5 TARGET DETECTION

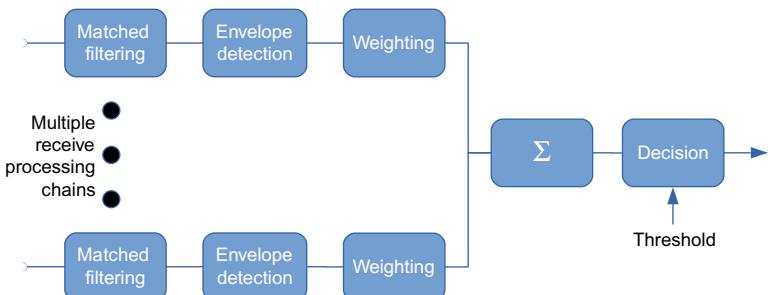
Highly centralized detection algorithms may fuse radio or video data prior to determining whether a target is present (i.e., through thresholding). A detector that combines radio signals, as shown in Fig. 5.2, strives to make gains through coherent summation of target returns. This requires the system to have knowledge of relative timing and position between all included transmitters/receivers such that the error of the knowledge is small compared with the transmitted wavelength. Even with a high degree of system coherency, any unknown delays or phase shifts introduced by the target (e.g., a complex unknown RCS) may preclude full coherent gains from being made at a hypothesized position.

In cases where coherent gains cannot be made in a centralized detector, fusion of video signals (i.e., following envelope detection), as shown in Fig. 5.3, can still

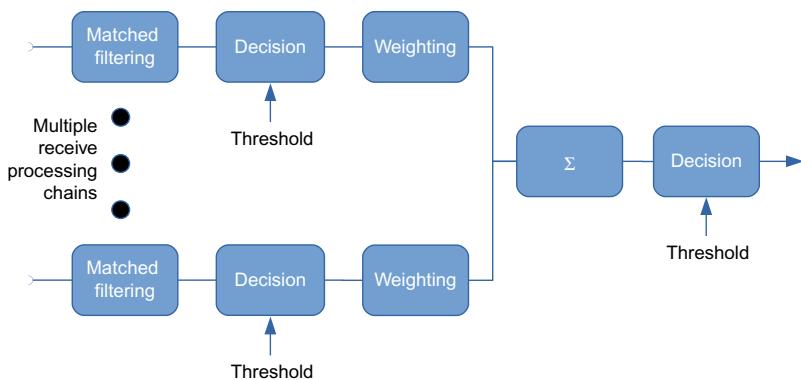


**FIG. 5.2**

Centralized coherent detector.

**FIG. 5.3**

Centralized noncoherent detector.

**FIG. 5.4**

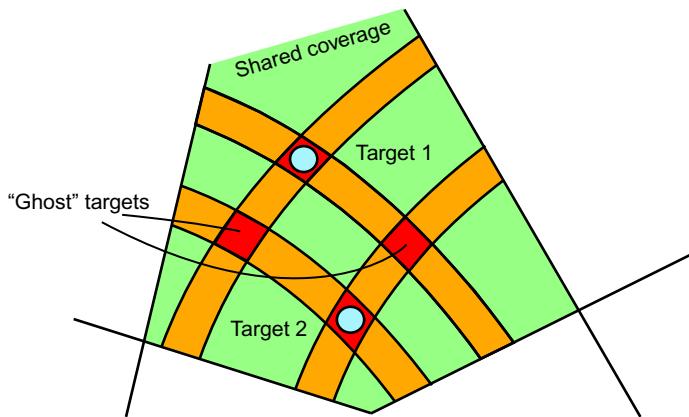
Decentralized detector.

provide detection performance improvements. Both types of centralized detectors apply weightings to maximize detection performance—these weightings may include estimates of amplitude and noise power. Full details of “optimal” detectors for detection performance can be found in [6].

A decentralized detector involves decision making prior to fusion; an example detector can be seen in Fig. 5.4. Some consideration must be given to how the nonlinear output of each individual decision is passed onward for fusion (e.g., Is peak detection used?).

## 5.6 TARGET RESOLUTION

At the fusion stage, detectors handling multiple targets have the potential for ambiguity. An example of the ambiguity is shown in Fig. 5.5 with the appearance of ghost targets. Similar to the problem of ghost targets is that caused by multipath, and/or

**FIG. 5.5**

Multistatic radar target association problem manifested as ghost targets.

jamming through retransmission of received signals—although these are not specifically a multistatic problem. The general solution to target association problems comes from the additional information provided by an increased number of transmitter-receiver pairs, and through use of Doppler information and by tracking over time.

Multistatic radar imaging represents a viable solution to the problem of multistatic target association. Furthermore, imaging may yield efficacy on the problem of interpreting low-resolution SAR/ISAR images that also suffer from shadowing.

## 5.7 TARGET LOCALIZATION

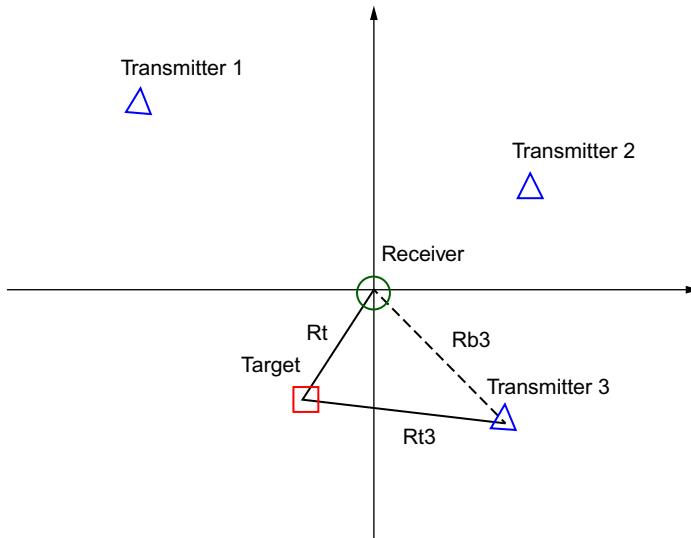
A multistatic radar permits three-dimensional target localization, if an adequate number of transmitters and/or receivers are present. The bistatic range obtained from such systems defines ellipsoids whose foci are at the position of the transmit-receive pair. At least three such ellipsoids are required to define a single target position where the ellipsoids intersect at a single point.

A multistatic radar in the discussion that follows is taken to be one comprising multiple transmitters with just one receiver. Note though that a system with one transmitter and multiple receivers could also be modeled in the same way.

The multistatic geometry is shown in Fig. 5.6. Much of what is presented here is based on Malanowski's work on localization using passive bistatic radar [20, 21].

The receiver is positioned at the origin  $[0, 0, 0]'$  of the configuration. There are  $N_{Tx}$  transmitters (in this example, three), with the  $i$ th transmitter located at  $[x_i, y_i, z_i]'$ . The target is located at  $[x_t, y_t, z_t]'$ . As the receiver is centered in the coordinate geometry at  $[0, 0, 0]'$ , the range from the receiver to the target is:

$$R_t = \sqrt{x_t^2 + y_t^2 + z_t^2} = \|\mathbf{x}_t\|, \quad (5.1)$$

**FIG. 5.6**

Geometry of a multistatic radar system.

where  $\|\mathbf{x}\| = \sqrt{\mathbf{x}' \mathbf{x}}$  is the norm of  $\|\mathbf{x}_i\|$ . The distance between the target and the  $i$ th transmitter is:

$$R_{ti} = \sqrt{(x_i - x_t)^2 + (y_i - y_t)^2 + (z_i - z_t)^2} = \|\mathbf{x}_i - \mathbf{x}_t\|. \quad (5.2)$$

The range from the receiver to the  $i$ th transmitter (the baseline range) is:

$$R_{bi} = \sqrt{x_i^2 + y_i^2 + z_i^2} = \|\mathbf{x}_i\|. \quad (5.3)$$

Multistatic passive radar measures the range sum from the transmitter to the target and then to the receiver and subtract the distance between the receiver and the transmitter. That is  $R_{ti} + R_t - R_{bi}$ . For convenience we use the transmitter-target-receiver range ( $R_{ti} + R_t$ ), which can easily be corrected by subtracting the known baseline range to the measured bistatic range:

$$R_{si} = (R_{ti} + R_t - R_{bi}) + R_{bi} = R_{ti} + R_t. \quad (5.4)$$

We define a vector of the transmitter-target-receiver ranges for each of the  $N_{Tx}$  transmitters:

$$\mathbf{r} = \begin{bmatrix} R_{s1} \\ R_{s2} \\ \vdots \\ R_{sN_{Tx}} \end{bmatrix}. \quad (5.5)$$

The goal is to obtain an estimate of the target position,  $\tilde{\mathbf{x}}_t = [\tilde{x}_t, \tilde{y}_t, \tilde{z}_t]'$ . From an estimated position, a vector of estimated ranges can be calculated:

$$\tilde{\mathbf{r}}(\mathbf{x}_t) = [\tilde{R}_{s1}, \tilde{R}_{s2}, \dots, \tilde{R}_{sN_{\text{TX}}}]'. \quad (5.6)$$

To localize the target we attempt to minimize the norm of the error between the measured ranges and the vector of ranges corresponding to the estimated position:

$$\hat{x}_t = \arg \min_{\tilde{x}_t} \|\mathbf{r} - \tilde{\mathbf{x}}_t\|. \quad (5.7)$$

The nonlinear relationship between the target location and its bistatic parameters results in a nontrivial optimization problem. In addition the cost function may have local minima, which means that standard optimization techniques may be ineffectual. They may also be computationally intensive and the target location may need to be estimated in real time. Therefore, a closed form solution is preferred.

Rearranging Eq. (5.4) in the following way:

$$R_{si} - \sqrt{x_i^2 + y_i^2 + z_i^2} = \sqrt{(x_i - x_t)^2 + (y_i - y_t)^2 + (z_i - z_t)^2}, \quad (5.8)$$

means that it is possible to approximately satisfy the optimization equation. Taking the square both sides and rearranging the terms yield the following equation:

$$x_i x_t + y_i y_t + z_i z_t - R_{si} \sqrt{x_i^2 + y_i^2 + z_i^2} = \frac{1}{2}(x_i^2 + y_i^2 + z_i^2 - R_{si}^2). \quad (5.9)$$

We introduce two substitutions to simply Eq. (5.9):

$$\mathbf{S} = \begin{bmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ \vdots & \vdots & \vdots \\ x_{N_{\text{Tx}}} & y_{N_{\text{Tx}}} & z_{N_{\text{Tx}}} \end{bmatrix} \quad (5.10)$$

and

$$\mathbf{z} = \frac{1}{2} \begin{bmatrix} x_1^2 + y_1^2 + z_1^2 - R_{s1}^2 \\ x_2^2 + y_2^2 + z_2^2 - R_{s2}^2 \\ \vdots \\ x_{N_{\text{Tx}}}^2 + y_{N_{\text{Tx}}}^2 + z_{N_{\text{Tx}}}^2 - R_{sN_{\text{Tx}}}^2 \end{bmatrix}. \quad (5.11)$$

Substituting into Eq. (5.9) for each of the  $N_{\text{Tx}}$  transmitters gives:

$$\mathbf{S}\mathbf{x}_t = \mathbf{r}R_t. \quad (5.12)$$

There are two unknowns in this equation, the target-receiver range  $R_t$  and the target position  $x_t$ . Assuming that  $R_t$  is known then the solution to this equation in the least squares sense is:

$$\hat{\mathbf{x}}_t = (\mathbf{S}'\mathbf{S})^{-1}\mathbf{S}'\mathbf{z} + (\mathbf{S}'\mathbf{S})^{-1}\mathbf{r}R_t. \quad (5.13)$$

The target range, of course, needs to be estimated before Eq. (5.13) can be solved for the target position. Malanowski in [20] comprehensively describes two methods of estimating target range, namely the Spherical Interpolation (SI) method and the

Spherical Intersection (SX) method. In the presence of measurement errors the SX method has been shown to produce more accurate results than the SI method. Derivations and example simulations are included in [20] and will not be reproduced here.

---

## 5.8 SYNCHRONIZATION CONSIDERATIONS FOR MULTISTATIC RADAR

Monostatic radars employ a single stable reference oscillator, from which all timing and frequency sources are derived [21]. This is a necessary consideration in order for the radar to be coherent, as received waveforms are “phase compared” to the transmitted reference signal. Any changes in phase during successive pulses provides useful information about the target, such as motion in the radial direction for Doppler processing.

Short-term frequency instabilities from the oscillator’s phase noise may hamper the performance of the radar by deteriorating the accuracy to which a phase comparison can be made. High phase noise references will mask slowly moving targets in the Doppler domain.

Multistatic radars, on the other hand, have nodes that are spatially separated, with baselines that may be in the order of kilometers. Using a single oscillator as a reference for all of the nodes is often practically challenging. For this reason, individual nodes of a multistatic radar can be designed with their own stable oscillators.

Clearly this introduces a challenge, as the phase noise present in each node’s oscillator will vary randomly. Depending on how stable the node oscillators are the ability to reliably measure any phase difference from the reference transmitter to the receiver may be affected.

Furthermore, the local clock derived from the reference at every node determines when each event in the timing diagram must occur. Unfortunately, no two clocks are identical and clocks will drift apart based on the noise contributions present in their oscillators. If the receiver does not report the same time as the transmitter, the order of events will be out by the time difference between the nodes. This will be most notable when a range measurement is performed.

As an example, if the receiver’s clock lags the transmitter’s by just 10 ns, the range measurement will include an additional 10 ns, which equates to the target being detected a further 3 m away (or 1.5 m when considering monostatic round-trip).

To avoid temporal and phase mismatches, the nodes are synchronized using one of several techniques. GPS Disciplined Oscillators (GPSDOs) are commonly used, as they provide a simple means of synchronizing nodes with relatively high accuracy (under 10 ns between nodes). A GPSDO uses the L1 carrier of GPS satellites to provide each node accurate UTC time, a PPS signal and a stable reference frequency, usually at 10 MHz. The device is phase-locked to the carrier with a PLL and thus traces the long-term stability of the satellites’ atomic clocks. The short-term stability of the device is reliant on the VCO used and its own phase noise characteristics.

GPSDOs are affordable in comparison with atomic clocks and offer similar stability performance over long periods of use. They are, however, reliant on GPS coverage which may at times fail and cause the device to enter a free-running oscillator mode. This also implies that the nodes may be prone to jamming.

Fiber-based Ethernet networks such as the White Rabbit network [10] developed by CERN are gaining traction in subnanosecond synchronization schemes. The system uses a variation of the Precision Time Protocol (PTP, IEEE1588) and Synchronous Ethernet to share a common clock from a highly stable grandmaster clock (atomic or GPS) to thousands of nodes. The time accuracy between any node and the grandmaster is guaranteed to be lower than 1 ns. Fiber or cable-based systems will, however, require infrastructure, and therefore rely on a stationary system.

Wireless communication links are possible, where the clock of a transmitting station is shared between the receivers. Using a line-of-sight communication link and a timing protocol based on time of flight messages, the clock from the transmitting time master can be given to the slave nodes. A variation of the White Rabbit protocol as discussed earlier has been achieved using COTS wireless links, with an average accuracy of 50 ns [22].

---

## 5.9 SYSTEM CASE STUDY: NetRAD/NeXtRAD

### 5.9.1 NetRAD

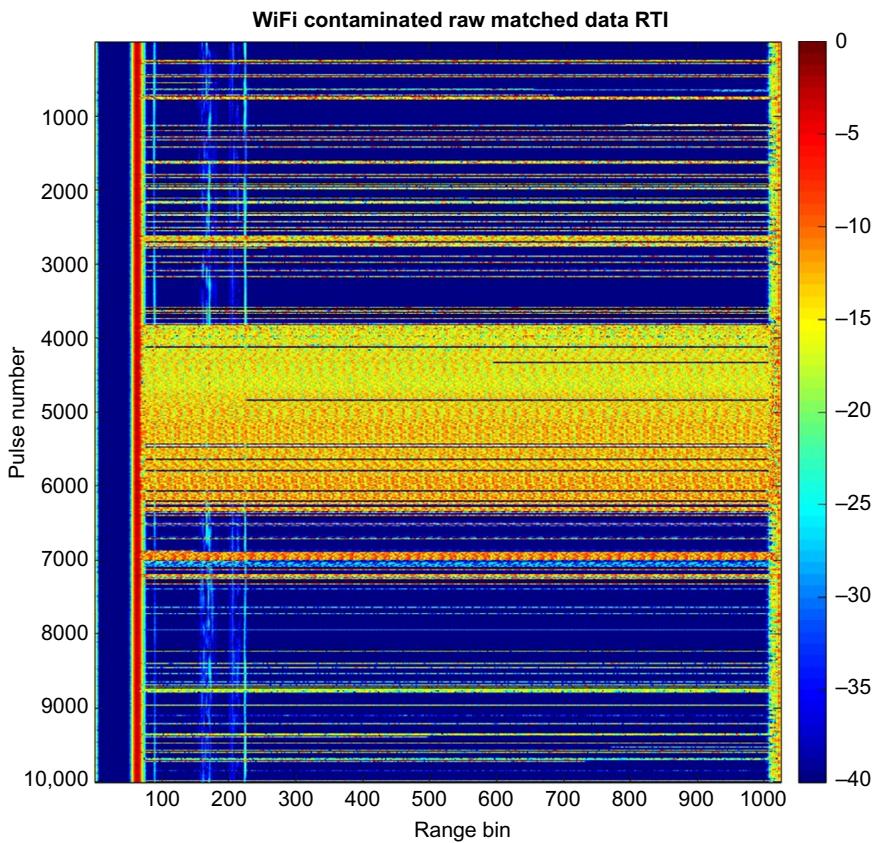
NetRad is a low-cost COTS multistatic radar, developed in collaboration between University College London [19] and the University of Cape Town, to explore the challenges associated with multistatic radar and to investigate what level of performance might be achievable. The addition of GPSDOs made an important contribution to the capability of NetRAD, that is, the ability to measure true multistatic clutter and targets [7, 8].

The system consists of three spatially diverse “nodes,” with each node capable of transmitting and receiving (both monostatically and bistatically) over a 50-MHz bandwidth at 2.4 GHz (S-Band). The initial aim of NetRAD was to achieve full spatial coherency through wired time synchronization and careful calibration over relatively short ranges. Antenna positions were fixed during operation—where typically wide beamwidth antennas were used to allow the widest area of shared coverage between nodes.

NetRAD has been deployed on numerous measurement campaigns, with the largest trial being sea clutter measurements in South Africa in 2010. Large amounts of littoral sea clutter data were collected during the 2010 trials. Extensive analysis of the data and how they compare with monostatic clutter has been performed in [13, 23].

As NetRAD operates in S-Band, it is highly susceptible to interference from Wi-Fi signals. An example of Wi-Fi interference corrupting NetRAD data is shown in Fig. 5.7.

In recent work, Jonkers in [24] has applied an LMS suppression algorithm iteratively across each range bin to suppress the frequency-hopping Wi-Fi interference,

**FIG. 5.7**

Heavy Wi-Fi contamination of NetRAD data.

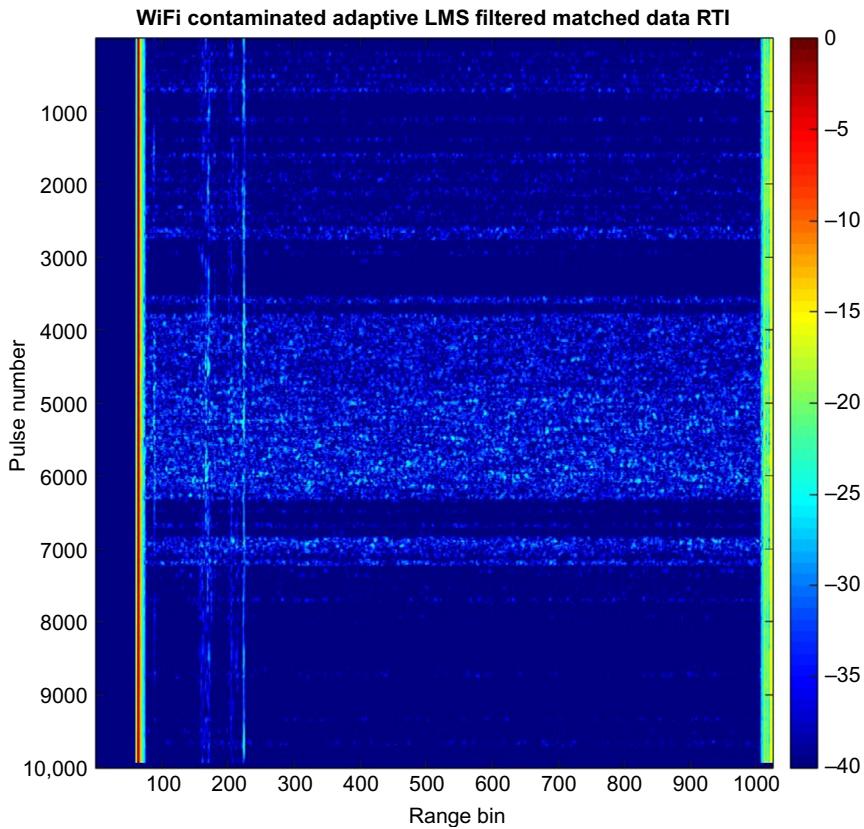
while adapting the algorithm to prevent it from suppressing the desired signal. An example of the NetRAD data after adaptive interference removal is shown in Fig. 5.8.

Jonkers also made use of the Julia programming language [25] to implement parallel pulse compression and pulse-Doppler processing algorithms for NetRAD (and NeXtRAD in future) using a multicore CPU.

NetRAD has also been used in multiple experimental measurements, the data from which was passed through various different detectors. These ranged from centralized coherent detection algorithms, focused on detection performance, to more decentralized methods that tended to focus more on the resolution of multiple targets.

One NetRAD experiment focused on the system's capability to detect and localize multiple moving targets. The measurement scenario is depicted in Fig. 5.9. The targets under observation in this experiment were two people walking through (0, 120).

An example of the NetRad plot extraction showing resolution of the aforementioned multiple targets (two people) is shown in Fig. 5.10. This result pertains to the

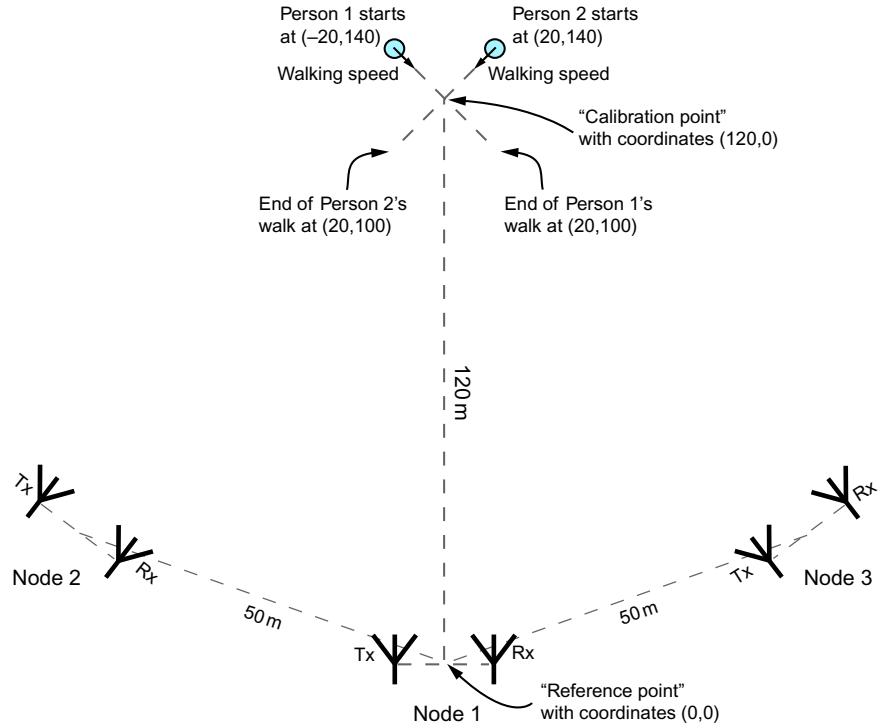
**FIG. 5.8**

Same NetRAD data as in Fig. 5.7, but here the Wi-Fi interference has been removed using an adaptive LMS algorithm.

case of decentralized detection where the system performed 30 detection measurements every 2 s. Each of the 30 measurements was used to produce a 2D position and velocity plot output (Fig. 5.10). Individual multistatic target position and velocity measurements are represented by arrows; where the base represents the position of the target and the length and direction of the arrow represents the target velocity.

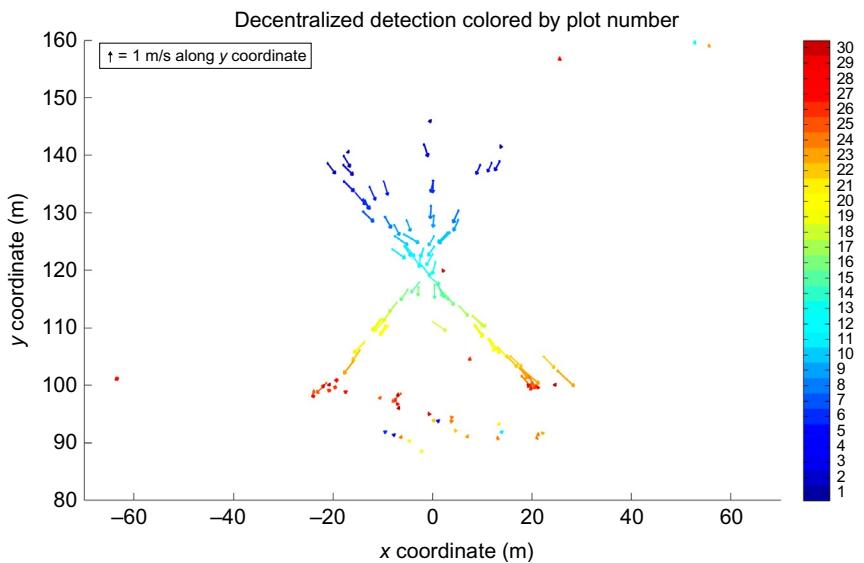
NetRAD was able to successfully detect and localize the two people consistently and tests were performed for different threshold values. The outlying arrows present in Fig. 5.10 represent false alarms as a result of imprecise thresholding. The lessons learned in this experiment have been useful in informing the design of NeXtRAD, particularly the aspects of NeXtRAD calibration, which will be described shortly.

A logical extension to the NetRAD experiment of detecting and locating human targets was to investigate the system's ability to discern the human micro-Doppler signature. Doughty in [12] describes the micro-Doppler experiments. We show in Fig. 5.11 a spectrogram of micro-Doppler shifts over a period of 0.5 s. The target



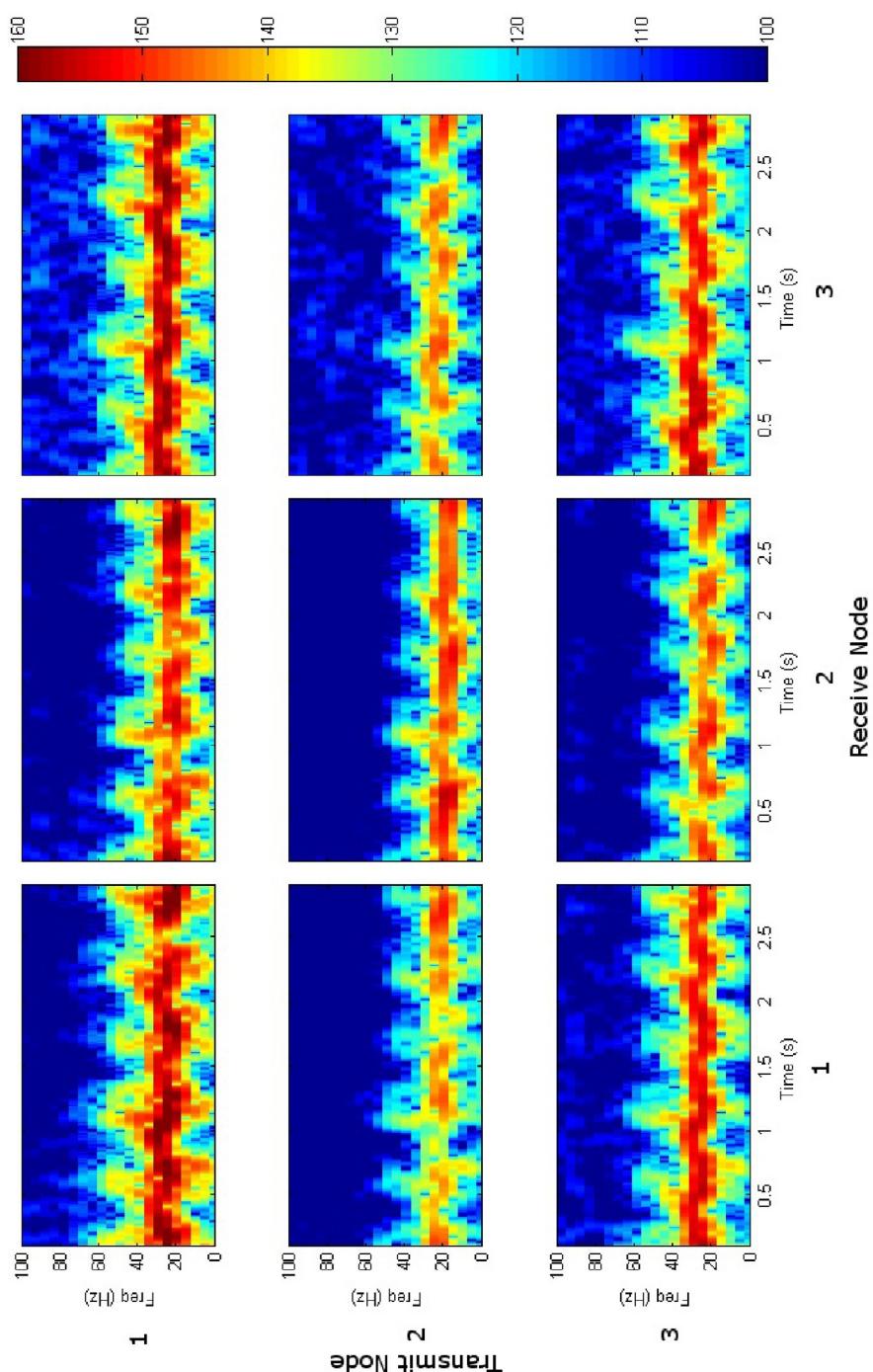
**FIG. 5.9**

NetRad test setup.



**FIG. 5.10**

Detection and localization of two people using NetRAD in the configuration depicted in Fig. 5.9.

**FIG. 5.11**

Micro-Doppler effect for a walking person.

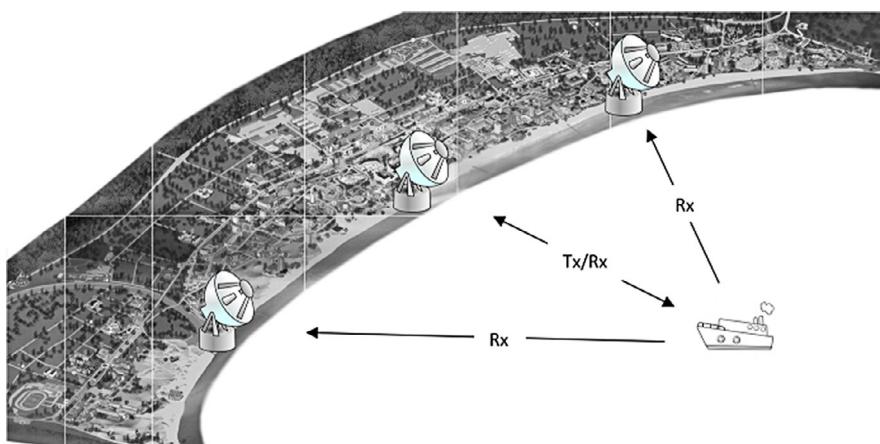
in this experiment is a human walking in the configuration described in Fig. 5.9. The oscillations correspond to a leg and/or arm swing.

The micro-Doppler effect was not investigated in greater detail in [12] because the 2.4 GHz carrier frequency used by NetRad was not well suited to resolving micro-Doppler modulations. This was a further reason why the next generation system, NeXtRAD, was designed to be dual frequency having both L-band and X-band operating options.

### 5.9.2 NeXtRAD

The NetRad system has evolved to incorporate research into multistatic, multiband networked radar and has resulted in the development of the NeXtRAD radar system. NeXtRAD is being developed under collaboration between University College London and the University of Cape Town with financial support from, inter alia, the US ONR-G and the South African National Research Foundation. NeXtRAD is a fully polarimetric multistatic radar that operates in both X- and L-band, and at the time of writing, the NeXtRAD team is preparing for their first measurement campaign along the coast of South Africa in late 2016 [18].

NeXtRAD has been designed to utilize two frequency bands as opposed to one, namely L-band (1.3 GHz) and X-band (8.5 GHz) and will also be fully polarimetric. Fig. 5.12 shows the basic node geometry of the NeXtRAD system. The radar consists of one active Tx/Rx node and two receive-only nodes. There is no wired-connection between nodes, which permits wider spatial separation between nodes than was possible with early iterations of NetRAD. The central active node of NeXtRAD will transmit either a horizontally or vertically polarized pulse in either X- or L-band,



**FIG. 5.12**

---

NeXtRAD antenna layout with central Tx/Rx node and two passive Rx-only flanking nodes.

while the target echo can be received by all three nodes in either horizontal or vertical polarization.

Multistatic, polarimetric, measurements of clutter and targets are very scarce. Obtaining quality measurement data and performing analysis is the first step toward building suitable prediction models for this new generation of networked, polarimetric radars that operate bistatically and multistatically. Polarimetry is well understood, especially in the field of imaging radar. However, most of the models used are based on monostatic SAR systems, so NeXtRAD will be an important testbed for theoretical modeling for bistatic and multistatic SAR.

Furthermore, a study by the NATO SET-164 Technical Team has indicated that some widely used propagation simulators such as ITU have poor ability to accurately predict received signal levels for low grazing angle bistatic radars. The technical team had measurement data from three different bistatic radars and then attempted to corroborate measurement with simulation. A portion of their findings has been published in [26]. It is evident, therefore, that systems like NeXtRAD will become increasingly valuable for improving the theoretical and practical understanding of bistatic and multistatic radar clutter and target detection performance and for improving the fidelity and reliability of radar modeling tools.

### 5.9.3 CALIBRATION OF MULTISTATIC POLARIMETRIC RADAR

The topic of multistatic, fully polarimetric, radar calibration is of paramount importance. It becomes essential to have high-fidelity discrimination between the H and V components. The antennas used must ensure that there is very precise alignment between the H and V feed probes, thereby ensuring orthogonality between the two planes of polarization. [Fig. 5.13](#) shows the NeXtRAD X-band antenna. It can



**FIG. 5.13**

Dual polarized X-band conical horn antenna.

**FIG. 5.14**

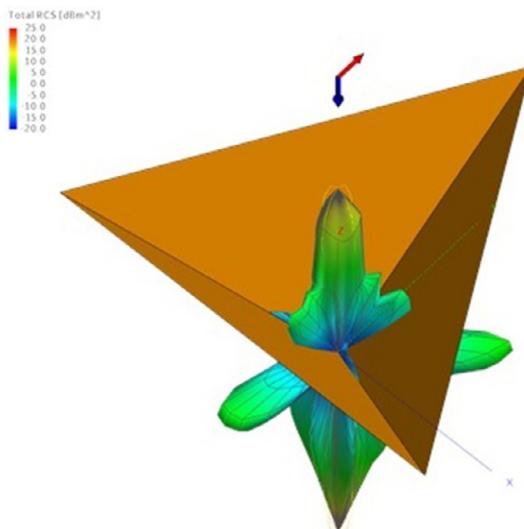
Dual polarized X-band conical horn antenna under test at UCT.

be seen that the conical horn antenna has been engineered to ensure close to perfect alignment between the mounting-bracket and the H-probe, while both are orthogonal to the V-probe. The same considerations apply in the L-band case.

[Fig. 5.14](#) shows the conical horns configured as part of the Tx/Rx monostatic node of NeXtRAD.

The antenna engineering process ensures polarization precision on both transmit and receive. The next step is to evaluate the radar with certain calibration targets so that the received response can be measured under the controlled conditions of an anechoic chamber. The reasons for performing measurements in an anechoic chamber are (i) because reflections are reduced, (ii) assumptions about how well the antenna has been engineered can be quantified, and (iii) the calibration targets used (e.g., trihedral) yield a deterministic response so that theory and measurement can be carefully matched. Once the chamber measurements have been performed, the process is repeated in the real environment. That is, the radar measures the response from calibration targets when operating in a complex scattering environment. The deviation in polarization response between the idealized case (anechoic chamber) and the real environment can be compensated to ensure that a high degree of orthogonality exists so that the scattering matrix for any given measurement has a minimum of cross-pol components.

Any passive device that can reflect microwave energy back to the radar can be used as a point of spatial reference. Passive corner reflectors are often used for geometric correction. However, alignment is critical to receive maximum energy [27]. Trihedral corner reflectors will be used for NeXtRAD's initial monostatic measurements. The collected data will be analyzed and used for calibration techniques to remove impurities in the data. Two commonly used trihedral corner reflectors are

**FIG. 5.15**

Triangular trihedral corner reflector with its RCS response simulated in FEKO.

the square and triangular reflectors, as shown in Fig. 5.15. The corner reflectors have been simulated in FEKO to observe the behavior of the reflected power.

The RCS of these corner reflectors is well known so that they can be used to calibrate the received power level. The RCS of a triangular trihedral is:

$$\sigma = \frac{4\pi a^4}{3 \lambda^2} \quad (5.14)$$

and RCS of a square trihedral is:

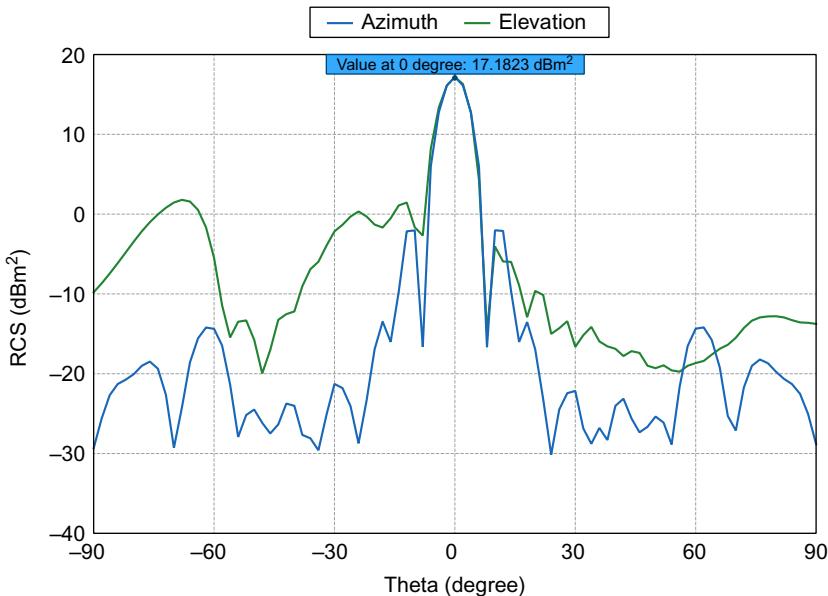
$$\sigma = 12\pi \frac{a^4}{\lambda^2}. \quad (5.15)$$

These equations can be used to find the RCS of the reflectors where  $\lambda$  is the freespace wavelength and  $a$  is the length of the corner reflector.

#### 5.9.4 CORNER REFLECTORS FEKO SIMULATION

The corner reflectors are configured in FEKO as follows:

- An EM planar wave with vertical polarization.
- Angle of incident is 45 degrees azimuth and 35 degrees elevation from the center of the corner reflector.
- For X-band, the wavelength  $\lambda$  is 35.29 mm.

**FIG. 5.16**

FEKO results for the triangular corner reflector with a length of  $10\lambda$ .

- The length  $a$  is simulated as  $10\lambda$  for both triangular and square trihedral corner reflectors.

After the configuration, the simulation extracts the radiation pattern and the results can be observed.

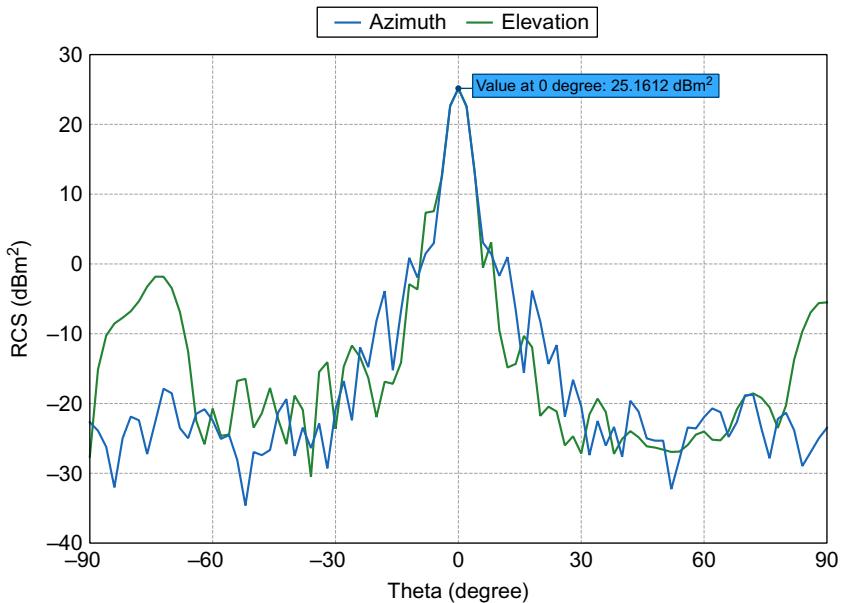
[Fig. 5.16](#) shows the radiation pattern of the triangular trihedral corner reflector in the azimuth and elevation planes.

[Fig. 5.17](#) shows the radiation pattern of the square trihedral corner reflector in the azimuth and elevation planes.

From observation, the triangular reflector has a wider main beam compared with the square reflector. However, the square reflector has a higher peak RCS of  $25.16 \text{ dBm}^2$  compared with the triangular reflector with  $17.18 \text{ dBm}^2$ .

It has been concluded that the triangular trihedral corner reflector will be more suitable as a calibration target for NeXtRAD measurements because they are physically smaller and yield a wider half-power response. The wider half-power response of the triangular reflector, the more leeway there will be for alignment between the radar and the calibration target.

Trihedral corner reflectors are currently being built as calibration targets for NeXtRAD. Measurements will be performed before the end of 2016. The results obtained are anticipated to be insightful for the general theme of multistatic, polarimetric, and radar calibration.

**FIG. 5.17**

FEKO results for the square corner reflector with a length of  $10\lambda$ .

## 5.10 CONCLUSIONS

This chapter has provided a general introduction to multistatic radar. Details of the most important system/performance characteristics of multistatic systems have been considered. In particular, we have discussed the possible information increase that a multistatic radar can yield in a given surveillance scene. Technological progress in SDR and FPGAs permits relatively fast prototyping of multistatic radar demonstrators—opening the possibility for a greater number of demonstrator systems to emerge in coming years.

We have provided discussion on important signal processing techniques pertinent to multistatic radar, such as Centralized and Decentralized detection processing. The field of multistatic radar node synchronization has emerged as an interesting and dynamic research area.

This chapter concluded with a case study of the NetRAD and NeXtRAD systems. NeXtRAD represents an exciting new generation of multifrequency, fully polarimetric, and multistatic radar. NeXtRAD will perform its first field-measurement campaign of sea clutter along the coast of South Africa in late 2016.

## ACKNOWLEDGMENTS

The authors owe a debt of gratitude to our colleagues, closest collaborators, and students. Particular thanks goes to Hugh Griffiths, Karl Woodbridge, Chris Baker, William Miceli, Matt Ritchie, Francesco Fioranelli, Shirley Coetzee, Waddah Al-Ashwal, Stephan Sandenbergh, Craig Tong, Andrew Van Der Byl, Rodolfo Azevedo Lima, Adrian Stevens, and the numerous post-docs and postgraduate students who are collectively known as the NeXtRAD consortium. To the students who have assisted with the preparations of this chapter for publication, we especially thank Po-Kai (Randy) Cheng, Simon Lewis, Thomas Düsterwald, Andrew Nicol, Stephanie Jonkers, and Stephen Paine.

We gratefully acknowledge Peralex Electronics, the SANDF for bursary funding over many years, and KACST for funding for S. Alhuweimal. The US ONR-G, FFI (Norway), IET AF Harvey Research Prize, NRF (South Africa), EPSRC, and the Royal Academy of Engineering. All funding bodies and partners have made substantial contributions to equipment, Postgraduate and Postdoctoral funds.

---

## REFERENCES

- [1] IEEE, IEEE Standard Radar Definitions 686-2008, IEEE, 22 pp.
- [2] N.J. Willis, *Bistatic Radar*, second ed., Technology Service Corporation, Silver Spring, MD, 1995.
- [3] K.S. Rani, T.K. Chaitanya, Detection of multiple targets by multistatic RADAR, *Int. J. Eng. Tech. Res.* 3 (7) (2015) 84–93.
- [4] J.M. Caspers, *Radar Handbook*, in: first ed., McGraw-Hill, New York, 1970. chap. 35.
- [5] E. Hanle, Survey of bistatic and multistatic radar, in: *IEE Proceedings F Communications, Radar and Signal Processing*, vol. 133, ISSN 0143–7070, 1986, pp. 587–595.
- [6] V.S. Chernyak, MIMO radars. What are they? in: *Proceedings of the 7th European Radar Conference*, 2010, pp. 137–140.
- [7] J.S. Sandenbergh, M.R. Inggs, W.A. Al-Ashwal, Evaluation of coherent netted radar carrier stability while synchronised with GPS-disciplined oscillators, in: *2011 IEEE RadarCon (RADAR)*, 1097-56592011, pp. 1100–1105.
- [8] J.S. Sandenbergh, M.R. Inggs, A common view GPSDO to synchronize netted radar, in: *2007 IET International Conference on Radar Systems*, 2007, ISSN 0537–9989, pp. 1–5.
- [9] M.R. Inggs, J.S. Sandenbergh, S.A.C. Lewis, Investigation of white rabbit for synchronization and timing of netted radar, in: *2015 IEEE Radar Conference*, 2015, pp. 214–217.
- [10] Open Hardware, ‘White Rabbit Overview’, 2016. [Online]. Available: <https://www.ohwr.org/projects/white-rabbit>. [Accessed: Oct 2016].
- [11] M. Weiss, Synchronisation of bistatic radar systems, in: *2004 IEEE International Geoscience and Remote Sensing Symposium, 2004. IGARSS ’04. Proceedings*, vol. 3, 2004, pp. 1750–1753.
- [12] S.R. Doughty, *Development and Performance Evaluation of a Multistatic Radar System* (Ph.D. Thesis), 2008.

- [13] H.D. Griffiths, W.A. Al-Ashwal, K.D. Ward, R.J.A. Tough, C.J. Baker, K. Woodbridge, Measurement and modelling of bistatic radar sea clutter, *IET Radar Sonar Navig.* 4 (2) (2010) 280–292.
- [14] D.W. O'Hagan, M. Ummenhofer, H. Kuschel, J. Heckenbach, A passive/active dual mode radar concept, in: 2013 14th International Radar Symposium (IRS), vol. 1, ISSN 2155–5745, 2013, pp. 136–142.
- [15] M. Inggs, C. Tong, D. O'Hagan, U. Böniger, U. Siegenthaler, C. Schuepbach, H. Pratisto, Noise jamming of a FM band commensal radar, in: 2015 IEEE Radar Conference, 2015, pp. 493–498.
- [16] R. Hosking, Putting FPGAs to Work in Software Radio Systems, sixth ed., Pentek Inc., Upper Saddle River, NJ, 2012.
- [17] J. Pendulum, RFNoC: RF network on chip, 2014 Cyberspectrum #12, 2014.
- [18] M. Inggs, H. Griffiths, F. Fioranelli, M. Ritchie, K. Woodbridge, Multistatic radar: system requirements and experimental validation, in: 2014 International Radar Conference, ISSN 1097–5764, 2014, pp. 1–6.
- [19] T.E. Derham, S. Doughty, K. Woodbridge, C.J. Baker, Design and evaluation of a low-cost multistatic netted radar system, *IET Radar Sonar Navig.* 1 (5) (2007) 362–368.
- [20] M. Malanowski, Signal Processing Algorithms and Target Localisation Methods for Continuous-Wave Passive Bistatic Radar, Habilitation, Warsaw University of Technology, Poland, 2012.
- [21] M.A. Richards, J.A. Scheer, W.A. Holm, Principles of Modern Radar, first ed., Scitech Publishing, Edison, NJ, 2010, pp. 417–454.
- [22] M. Rico, J.P. Aubry, C. Botteron, P.A. Farine, Ns-level time transfer over a microwave link using the PTP-WR protocol, in: 2015 Joint Conference of the IEEE International Frequency Control Symposium the European Frequency and Time Forum, ISSN 2327–1914, 2015, pp. 690–695.
- [23] W.A. Al-Ashwal, Measurement and Modelling of Bistatic Sea Clutter (Ph.D. Thesis), (2011).
- [24] S.C. Jonkers, Software Infrastructure for NeXtRAD Development in Julia Programming Language (Masters Dissertation), 2016.
- [25] Julia, ‘Julia High-Performance JIT Compiler’, 2016. [Online]. Available: <http://julialang.org/>. [Accessed: Oct 2016].
- [26] M. Malanowski, R. Haugen, M.S. Greco, D.W. O'Hagan, R. Plsek, A. Bernard, Land and sea clutter from FM-based passive bistatic radars, *IET Radar Sonar Navig.* 8 (2) 2014 160–166.
- [27] J.A. Richards, Remote Sensing With Imaging Radar, Springer-Verlag, Berlin, Heidelberg, 2009.

# Sparsity-based radar technique

# 6

Matthias Weiß

Fraunhofer FHR, Passive Radar and Anti-Jamming Techniques (PSR), Wachtberg, Germany

## 6.1 INTRODUCTION

These days many remote sensing and surveillance radar systems take advantage of technology improvements by employing a large frequency bandwidth and/or by increasing the number of transmit/receive nodes to obtain a highly flexible and adaptable system with high spatial resolution. In particular in MIMO radar systems with a huge instantaneous frequency bandwidth and a high number of transmit/receive nodes a vast amount of data is generated, which has to be processed. On the other hand the observed scene contains only a small number of targets which contribute to the received signal. Often it seems that there is an imbalance between the observed surveillance area, which is sparsely populated, for instance with a countable number of aircraft in the wide sky, an expensive fully digitized radar system with its huge amount of antenna elements generating a vast data stream, and the final output after the signal processing stage, which again might be very sparse as only the parameters of relative few objects of interest will be extracted.

Traditional signal processing methods have to process all data in order to estimate the parameters of these few targets. Many attempts have been made to reduce the required data rate for these fully digitized radar systems. For instance *Sparse Arrays* have been investigated to overcome the Nyquist criterion for spatial sampling [1, 2]. Another approach to reduce cost and hardware complexity is the *multiinput multi-output* (MIMO) radar [3]. Here  $N_{tx}$  independent transmit nodes and  $N_{rx}$  receive nodes yields  $N_{tx} \times N_{rx}$  different transmit-target-receive signals which increase the degrees of freedom. This can lead to an improvement in phase and angular estimation of the received signal. If transmit and receive nodes are placed around the common observation scene the obtained range resolution of the system is improved. However, these distributed systems produce a massive data stream which has to be processed by the signal processing chain. Due to strong sidelobes, provoked by the limited system bandwidth, nearby objects of interest can be concealed. These points raise the

question whether sparse signal processing techniques might help to overcome these problems and if they are applicable to radar.

With the beginning of the 21th century sparse signal processing, also noted as compressive sensing (CS), Refs. [4-10] have emerged, which inaugurated a new way of looking at the sensing/sampling paradigm. It represents a revolution in signal processing and sensor systems for collecting and processing data from sparse scenes. The primary intention of CS is to invert or reconstruct a  $K$ -sparse signal  $\mathbf{x}$  from a set of *linear* measurements  $\mathbf{y}$  in the form  $\mathbf{y} = \mathbf{Ax} + n$ , where  $n$  represents the noisy part, which is in real systems always present. The sparse reconstruction approach enables us to solve this linear problem even in the underdetermined case where the number of samples in space, time, and frequency is less than demanded by the Nyquist-Shannon criteria. If we can describe the scene/signal by a vector  $\mathbf{s}$  of dimension  $N$  with a  $K$ -sparse representation ( $\|\mathbf{x}\|_0 = K \ll N$ , with  $\|\mathbf{x}\|_0 = \sum_i x_i^0$  the modified  $\ell_0$ -norm returning the number of nonzero coefficients) in some orthonormal basis (e.g., wavelet, Fourier) or tight frame (e.g., curvelet, Gabor)  $\Psi$ , as  $\mathbf{x} = \Psi\mathbf{s}$ , and that this vector is compressible, which means that the vector is composed of only a few large coefficients and other with small values ( $\|\mathbf{x} - \mathbf{x}(K)\|_2$  decreases quickly to zero with growing  $K$ ), sparse reconstruction techniques are able to recover the sparse signal  $\mathbf{x}$  with a very high probability only using few measurements  $M = \mathcal{O}(K \log(N/K))$  of  $\mathbf{y}$  are necessary to solve the convex  $\ell_1$  optimization problem:

$$\min \|\mathbf{x}\|_1, \text{ subject to } \|\mathbf{y} - \mathbf{Ax}\|_2 < \sigma, \quad (6.1)$$

where  $\sigma \geq 0$  is a threshold parameter for systems where noise is present and has to be determined by well-established statistical methods. There exist some well-known algorithms to solve this minimization problem: Basis Pursuit Denoising (BPD) [11], the Orthogonal Matching Pursuit (OMP) [12], the Compressive Sampling Matched Pursuit (CoSaMP) [13], and the SPGL1 [14] algorithms.

Owing to these attractive properties sparse recovering techniques found much attraction in the field of radar with its sparse scenes over the past years. One of the earliest papers on CS applied to radar is from Baraniuk and Steeghs [15]. Nowadays there are many research projects with the focus of understanding the limitations and constraints of sparse reconstruction techniques applied to high resolution radar, interferometric SAR, inverse SAR, Moving Target Indication (MTI), and direction-of-arrival (DOA) estimation.

It has been shown that CS techniques provide a guaranteed stable solution of the reconstructed sparse signal  $\mathbf{x}$  for a given sensing matrix  $\mathbf{A}$  if it satisfies the restricted isometry property (RIP), as introduced by Candès and Tao [16].

An  $M \times N$  matrix  $\mathbf{A}$  satisfies the RIP of order  $K$  if there exists a  $\delta_K \in (0, 1)$  such that [17-22]:

$$(1 - \delta_K) \|\mathbf{x}\|_2^2 \leq \|\mathbf{Ax}\|_2^2 \leq (1 + \delta_K) \|\mathbf{x}\|_2^2 \quad (6.2)$$

for all  $K$ -sparse vectors  $\mathbf{x} \in \mathbb{C}^N$ . When  $\delta_K$  is less than 1 this RIP implies that all of the submatrices of  $\mathbf{A}$  with  $K$ -columns are well-conditioned and close to an isometry.

If  $\delta_K \ll 1$  then there is a high probability to reconstruct the  $K$  sparse signal  $x$  with the sensing matrix  $A$ . It has been shown that a random selection of the measurement matrix  $\Phi$  the RIP can be achieved with high probability [20, 21], however this cannot always be applied to real radar systems and scenes.

The RIP criterion provides a guarantee for the recovery of a  $K$ -sparse signal, however, any of these properties is hard to verify for a general sensing matrix  $A$  as  $\binom{N}{K}$  submatrices have to be considered during the computation. In many cases it is preferred to determine a characteristic of  $A$  that is much easier to compute and delivers a practical recovery guarantee. Such a property is the column coherence or mutual coherence of matrix  $A$  [23–25]. It is defined as the maximal absolute value of the cross-correlations between the  $N$  columns of matrix  $A \in \mathbb{C}^{M \times N}$ . This column coherence should not be mixed up with the coherence of the sensor system. Formally, let  $a_1, \dots, a_N$  be the columns of the matrix  $A$ . The coherence of  $A$  is then defined as:

$$\mu(A) = \max_{1 \leq i \neq j \leq N} \frac{|\langle a_i^*, a_j \rangle|}{\|a_i\|_2 \|a_j\|_2} \quad (6.3)$$

with  $|\langle \cdot, \cdot \rangle|$  the inner product between any two columns  $a_i, a_j$  and  $1 \leq i \neq j \leq N$ .

Another way to determine the column coherence of  $A$  with normalized columns  $\left(\frac{a_i}{\|a_i\|_2}, 1 < i < N\right)$  is first to calculate the Gram matrix  $(A^H A)$  and then subtract the identity matrix and find the maximal entry of this matrix:

$$\mu(A) = \max a_{ij} \in (A^H A - I). \quad (6.4)$$

Calculating the column coherence by Eq. (6.4) has the advantage that very efficient computer algorithms exist. It can be shown that the column coherence  $\mu(A)$  is always in the range of  $\mu(A) \in \left[\sqrt{\frac{N-M}{M(N-1)}}, 1\right]$ . The lower bound is also known as the Welch bound and is for  $N \ll M$  approximately  $\mu(A) = 1/\sqrt{M}$  [26, 27].

In the case of  $\mu(A) = 1$  there are at least two columns aligned. This represents the worst case scenario: maximum coherence. In the other extreme, if  $\mu(A) = \sqrt{\frac{N-M}{M(N-1)}}$ , the best scenario exists: maximal incoherence. If  $A$  is a square matrix and the columns are orthonormal  $\mu(A) = 0$ . For a good convergence of any sparse recovery algorithms the column coherence  $\mu$  of the sensing matrix  $\mu(A)$  should be always  $< 1$ .

This concept can also be applied to nonsparse scenes if they can be transformed to a basis with a sparse coefficient vector. The measurements of the surveillance area  $y$  can be described by the following sensing matrix  $\Phi$  and the vector  $\sigma$ , which holds the description of the nonsparse scene:

$$y = \Phi \sigma \quad \Phi \in \mathbb{C}^{M \times N}. \quad (6.5)$$

If we can transform the nonsparse vector  $\sigma$  with the following transformation matrix  $\Psi$  into a new basis with a sparse vector  $s$ :

$$\sigma = \sum_{n=1}^N s_n \Psi = \Psi s. \quad (6.6)$$

This can be for instance the Fourier transformation. With this relation we obtain for Eq. (6.5) again a sparse description of the scene:

$$\mathbf{y} = \Phi \boldsymbol{\sigma} = \Phi \Psi \mathbf{s} = \mathbf{A} \mathbf{s} \quad \text{with } \mathbf{A} = \Phi \Psi \in \mathbb{C}^{M \times N}. \quad (6.7)$$

The mutual coherence between two orthonormal bases  $\Phi = (\phi_1, \dots, \phi_M)$  and  $\Psi = (\psi_1, \dots, \psi_M)$  of  $\mathbb{C}^M$  is given by:

$$\mu(\Phi, \Psi) := \sqrt{M} \max_{1 \leq i, j \leq M} \frac{|\langle \phi_i, \psi_j \rangle|}{\|\phi_i\|_2 \|\psi_j\|_2}. \quad (6.8)$$

The mutual coherence between the orthonormal bases  $\Phi$  and  $\Psi$  is of course closely related to the coherence of the system obtained by concatenation of the bases  $\Phi$  and  $\Psi$

$$\mu(\Phi, \Psi) = \sqrt{M} \mu(\mathbf{A}). \quad (6.9)$$

A small mutual coherence function means, that all submatrices of  $\mathbf{A}$  with maximum  $M$  columns are well-conditioned.

This paper provides an insight into some possibilities of sparse recovering techniques in the field of sub-Nyquist sampling in time, space, frequency and its possibility to overcome the limitation of the matched filter. Another part deals with the aspects of data fusion of sensor networks by employing a group sparsity approach developed in the framework of CS. One of the main algorithms developed in this field is the conditioned minimization of the  $\ell_1$ -norm of the vector describing the amplitude distribution of the scene under the condition that the measurements are compatible with the signal model. This theory is applicable for temporal as well as for spatial sampling.

The remaining paper is organized as follows. In Section 6.2 the focus of sparse reconstruction is in the temporal domain and its implementation as pulse-compression stage in fast and slow-time. Section 6.3 deals with the topic of recovering missing spectral information from a corrupted signal. Section 6.4 discusses the sparse reconstruction technique in the area of spatial sparsity and in Section 6.5 this technique is applied to combine the information from a bunch of sensors observing the same scene. Conclusion and indications for future development are covered in Section 6.6.

## 6.2 TEMPORAL SPARSITY

Let's assume that a simple pulse Doppler radar emits a periodic series of frequency modulated pulses. The reflected signals are received, downconverted, and digitized. Then digital signal processing algorithms extract the information from the almost continuous data stream. Conventional radar systems sample and process the received signals according to the Nyquist-Shannon criteria. This requires high-speed analog-digital converters (ADC) and ultra-fast digital processing stages to realize a continuously running matched filter (MF). Due to this, the design of a high resolution radar

system, with frequency bandwidth of up to hundreds of MHz or even GHz, is limited by the availability of high-speed components and in many cases the required technology is beyond the state-of-art. It might be also the case that the budget is the limiting factor as the technique is too expensive.

One approach to relax the hardware requirements of a pulse Doppler radar is by applying sub-Nyquist sampling and sparse recovering techniques, along with special designed receiver front-end [15, 28]. This method changes the hardware design of the radar system by a reduction of the demanded bandwidth of the ADC and the subsequent digital components. Along the signal processing chain the conventional matched filter is replaced by sparse signal recovering technique. This leads to a data stream much less than demanded by Nyquist-Shannon from a sparse scene populated with a few targets.

### 6.2.1 SPARSE SAMPLING IN RANGE

For simplicity let's assume that a radar illuminates the surveillance area with the waveform  $x_t(t)$ . Likewise there are only  $K$  targets presents and we can neglect clutter. With these assumptions the received signal  $y_t(t)$  is the sum of all echoes reflected by these targets. They are located at the distance  $r_k$ ,  $k = 1, \dots, K$ , which correspond to a time-delay of  $\tau_k = 2 r_k/c$  and the amplitudes of the received echoes  $s_k$  are proportional to the radar cross sections. The received signal is sampled and digitized at  $t = t_1, \dots, t_M$ . Hence, for sampling time  $t_m$  we can describe the signal by:

$$y_m = \sum_{k=1}^K s_k x_t(t_m - \tau_k). \quad (6.10)$$

In general the measurement vector  $\mathbf{y}$  is defined by:

$$\mathbf{y} = [y_1, \dots, y_M]^T = \mathbf{A}\mathbf{s}, \quad (6.11)$$

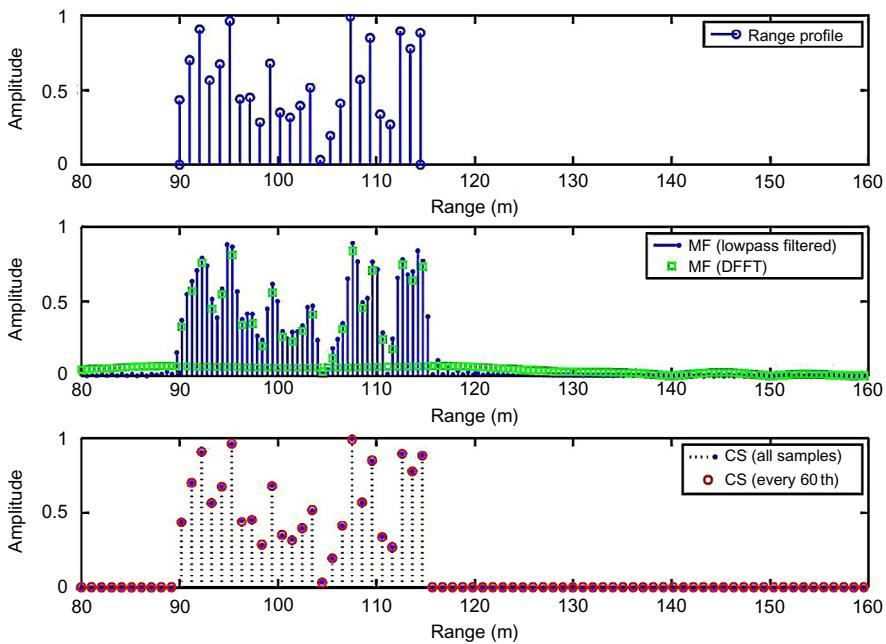
where vector  $\mathbf{s}$  contains the amplitudes  $s_n$  of all considered echoes  $\mathbf{s} = [s_1, \dots, s_N]^T \in \mathbb{C}^{N \times 1}$ . Each column of the sensing matrix  $\mathbf{A}$  is a description of what will be observed if a target is present at the corresponding scene point with echo amplitude of one. According to this column  $\mathbf{a}_n$  of the sensing matrix is just a time-delayed version of the transmitted signal:

$$\begin{aligned} \mathbf{a}_n &= [x_t(t_1 - \tau_n), \dots, x_t(t_M - \tau_n)]^T \\ \mathbf{A} &= \text{vec}(\mathbf{a}_1, \dots, \mathbf{a}_N) \in \mathbb{C}^{M \times N}. \end{aligned} \quad (6.12)$$

As we know that the scene is sparse populated we can estimate the number of targets and their reflection coefficients by solving the underdetermined equation set via the convex  $\ell_1$  optimization problem:

$$\min_{\mathbf{s}} \|\mathbf{s}\|_1 \text{ subject to } \|\mathbf{A}\mathbf{s} - \mathbf{y}\|_2 < \sigma. \quad (6.13)$$

A simulation was carried out with a radar emitting a pulse modulated by a linear frequency sweep with a bandwidth of 147 MHz. The sampling rate was set to twice the

**FIG. 6.1**

Pulse compression of a chirped pulse in the time domain by a matched filter approach and applying sparse reconstruction techniques.

bandwidth of the transmitted pulse, as required by the Nyquist-Shannon criteria. The results obtained by matched filter and by sparse reconstruction technique are shown in Fig. 6.1.

The top diagram shows the predetermined range profile which consists of 25 point scatterers. The result obtained by the matched filter with a sample rate of 300 MHz is shown in the middle diagram by the data represented by squares. To obtain the correct range representation the output of the matched filter has to be interpolated by sinc-functions, which is equal to a lowpass filtering of the signal and is represented by the circles with stems (blue). The bottom diagram shows the obtained results from the sparse reconstruction technique. The stems (blue) indicates the estimated range profile obtained when all measurement data are used. However, the advantage of this new reconstruction technique reveals when the sample rate is much less than the Nyquist rate. This is shown by the circles (red), which represents the range profile obtained from a measurement with a sample rate 60 times less than Nyquist. Under these conditions the matched filter stage fails to estimate any range profile.

### 6.2.2 SPARSE SAMPLING IN RANGE AND DOPPLER

For the second example we consider a pulse radar which emits a pulse train with a given pulse repetition time interval (PRI) or reciprocal pulse repetition frequency (PRF), respectively. The samples are then arranged into a matrix form with one index corresponding to the range (fast-time) and the other to the pulse index (slow-time). The Doppler frequency of the moving object can then be obtained by a FFT along the slow-time. As shown in the previous example sparse signal processing techniques allows us to implement a nonuniform or so-called co-prime sampling in range (fast-time). In the second example this technique is expanded to include also a co-prime sampling in slow-time to further reduce the amount of data which has to be processed. It will be shown that the estimation accuracy of the Doppler frequency of the moving target will not degrade.

Let an ordinary pulse Doppler radar emit  $M$  pulses with a pulse repetition interval (PRI) of  $\tau_{\square}$ . A single transmitted pulse is described by its baseband function  $x(t)$  and its continuous-time Fourier transformation  $X(\omega) = \int_{-\infty}^{\infty} x(t) e^{-j\omega t} dt$ . The frequency bandwidth of the pulse is  $B$  and we presume that  $X(\omega)$  has almost no energy at frequencies outside of  $B$ . With these constraints the transmitted signal  $x_{trans}(t)$  can be written in a matrix form as:

$$\begin{aligned} \mathbf{X}_{trans}(t) &= [\mathbf{x}_1^T(t), \dots, \mathbf{x}_M^T(t)] \\ \text{with } \mathbf{x}_m(t) &= x(t - (m-1)\tau_{\square}), \quad (m-1) \leq t \leq m\tau_{\square}. \end{aligned} \tag{6.14}$$

The matrix  $\mathbf{X}_{trans}$  consists of  $M$  columns which represent the equally spaced pulses  $x(t)$  with a pulse to pulse delay of  $\tau_{\square}$ . The range and Doppler resolution obtained by standard radar signal processing techniques is reciprocally proportional to the signal bandwidth  $\Delta r \sim 1/b$  and to the coherent processing time  $\Delta f_D \sim 1/M\tau_{\square}$ , respectively.

For the sake of simplicity let us assume that a scene contains  $K$  constant moving point-like targets. According to that the transmitted pulses  $x(t - (m-1)\tau_{\square})$  are reflected by these objects and propagate back to the receiver. Therefore the received signal is described by:

$$\begin{aligned} \mathbf{Y}_{received}(t) &= [\mathbf{y}_1^T(t), \dots, \mathbf{y}_M^T(t)] \\ \text{with } \mathbf{y}_m(t) &= \sum_{k=1}^K s_k x(t - \tau_k - (m-1)\tau_{\square}) e^{-j f_{Dk} (m-1)\tau_{\square}}, \quad (m-1) \leq t \leq m\tau_{\square}, \end{aligned} \tag{6.15}$$

where  $s_k$  is the complex amplitude corresponding to  $k$ th target radar cross section and the propagation attenuation,  $\tau_k = 2 r_k/c_0$  the time delay, and  $f_{Dk}$  the Doppler radial frequency, proportional to the radial velocity of the target.

For a co-prime sampling in slow-time we introduce the variable  $P \subset [1, \dots, M]$ , where the index  $p$  denotes the  $p$ th pulse transmitted at time  $m_p\tau_{\square}$ :

$$\mathbf{Y}_p(t) = [\mathbf{y}_1^T(t), \dots, \mathbf{y}_P^T(t)]. \tag{6.16}$$

By introducing index  $p$  we extend the sparse recovering technique in such a way that it can work likewise with sparse sampling in slow-time (pulse repetition or Doppler-domain). The objective of the following steps are to estimate accurately the target range and Doppler frequency, actually time delay  $\tau_l$  and Doppler shifts  $v_l$ , from the received signals  $\mathbf{y}_p(t)$ . It will be shown that again sparse recovering technique is able to recover this information with less samples in fast-time (range) and slow-time (quantity of transmitted pulses  $\text{supp}(P) < M$ ) without degrading the accuracy of the estimation. However, this will only be true if the observation time  $M\tau_{\square}$  is equal.

At first we transform the time-domain representation of the aligned received pulses  $\mathbf{Y}_{\text{received}}(t)$  from Eq. (6.15) into the Fourier domain with  $N$  discrete frequencies  $f_{Dn} \in (-f_D, \dots, f_D)$ :

$$\mathbf{Y}_p[n] = \frac{1}{\tau_{\square}} X[n] \sum_{l=1}^L s_l \cdot e^{-j 2 \pi f_{Dn} \tau_l} \cdot e^{-j v_l (m_p - 1) \tau_{\square}/\lambda} \quad (6.17)$$

with  $\lambda$  the wavelength of the center frequency of the transmitted modulated pulse and  $L$  the number of grid points in the time-delay/Doppler plane ( $s_l, v_l$ ). For the continuous case the sampling rate in fast-time and slow-time is determined by the pulse bandwidth ( $t_n \geq 1/2 B$ ) and by the demanded Doppler resolution ( $\sim \text{PRF}/\Delta f_D$ ), respectively. The unknown parameters of the targets ( $s_l, \tau_l, v_l$ ) are contained in the Fourier coefficients  $\mathbf{Y}_p[n]$ .

The measurement matrix  $\mathbf{Y} \in \mathbb{C}^{N \times P}$  with its  $p$ th column defined by  $\mathbf{y}_p[n]$  describes the measurement in fast- and slow-time in the Fourier domain and can be expressed by:

$$\mathbf{Y} = \mathbf{X} \mathbf{F}_{\tau} \mathbf{S} (\mathbf{F}_D)^T \quad (6.18)$$

with  $X[n]$  the Fourier transformed pulse at  $N$  discrete frequencies ( $f_{Dn} = -f_D, \dots, f_D$ ),  $\mathbf{X} = \frac{1}{\tau_{\square}} \text{diag}(X[n])$ ,  $\mathbf{F}_{\tau}$  the Fourier time shift matrix with its  $M$  columns and  $N$  selected Fourier coefficients ( $F_{\tau}^K(m, n) = \exp(-j 2 \pi f_{Dn} \tau_m)$ ), and  $\mathbf{F}_D \in \mathbb{C}^{N \times P}$  the Fourier Doppler shift matrix with its  $P$  rows ( $F_D^P = \exp(-j 2 \pi f_D (m_p - 1) \tau_{\square}/\lambda)$ ). The matrix  $\mathbf{S}$  of size  $M \times N$  holds the values  $s_l$  at the searching grid points  $L = M N$  in the range/Doppler plane ( $\tau_l, v_l$ ). As the scene contains only a limited number of targets the relation  $K \ll L$  is true and consequently  $\mathbf{S}$  is a sparse matrix with only  $K$  nonzero elements.

To estimate the coefficients of the sparse matrix  $\mathbf{S}$  we first consider that

$$\mathbf{Z} = \mathbf{X}^{-1} \mathbf{Y} \quad (6.19)$$

and then we obtain

$$\mathbf{Z} = \mathbf{F}_{\tau} \mathbf{S} (\mathbf{F}_D)^T. \quad (6.20)$$

For a successful parameter estimation of the sparse matrix elements  $s_{n, p}$  it has been shown for the noiseless case that for  $K$  targets the minimal number of samples of  $4K^2$  are needed, with  $N \geq 2 K$  (fast-time) and  $P \geq 2 K$  (slow-time), respectively [29, 30].

For the noise free environment we have to solve the following optimization problem to recover the nonzero elements of the sparse matrix  $\mathbf{S}$  [31]:

$$\min_{\mathbf{S}} \|\mathbf{S}\|_1 \text{ subject to } \mathbf{F}_\tau \mathbf{S} (\mathbf{F}_D)^T = \mathbf{Z}, \quad (6.21)$$

where  $\|\mathbf{S}\|_1 = \sum_{i,j} |S_{ij}|$  is the  $\ell_1$ -norm of  $\text{vec}(\mathbf{S})$ , where  $\text{vec}(\mathbf{S})$  vectorize the matrix  $\mathbf{S}$  by stacking the columns into a vector.

If there is any noise present Eq. (6.21) has to be adapted in such a manner that a threshold is taken into account:

$$\min_{\mathbf{S}} \left\{ \frac{1}{2} \|\mathbf{Z} - \mathbf{F}_\tau \mathbf{S} (\mathbf{F}_D)^T\|_2^2 + \lambda \|\mathbf{S}\|_1 \right\}, \quad (6.22)$$

where  $\lambda$  is a regularization parameter and  $\|X\|_2 = \sum_{i,j} |X_{ij}|^2$  is the  $\ell_2$ -norm of  $\text{vec}(X)$

To solve the underdetermined linear equation set defined in Eq. (6.22) several sparse recovering algorithms exist (comp. in [31]). An alternative procedure to estimate the coefficients of the sparse state matrix  $\mathbf{S}$  is to transform Eq. (6.21) into a vector version by using the following relation:

$$\text{vec}(\mathbf{F}_\tau \mathbf{S} \mathbf{F}_D) = (\mathbf{F}_D^T \otimes \mathbf{F}_\tau) \text{vec}(\mathbf{S}), \quad (6.23)$$

where  $\otimes$  is the Kronecker operator and  $\text{vec}(\mathbf{B})$  is the operator which stacks the columns of a matrix  $\mathbf{B}$  into a vector  $\mathbf{b}$ .

Applying Eq. (6.23) to Eq. (6.21) yield:

$$\mathbf{y} = \mathbf{C} \mathbf{s}, \quad (6.24)$$

where  $\mathbf{y}$  is the stacked version of matrix  $\mathbf{Z}$ ,  $\mathbf{C} = \mathbf{F}_D^T \otimes \mathbf{F}_\tau$ , and  $\mathbf{s} = \text{vec}(\mathbf{S})$ . With this equation we can find the sparse solution of  $\mathbf{s}$  by solving the  $\ell_1$ -norm minimization problem:

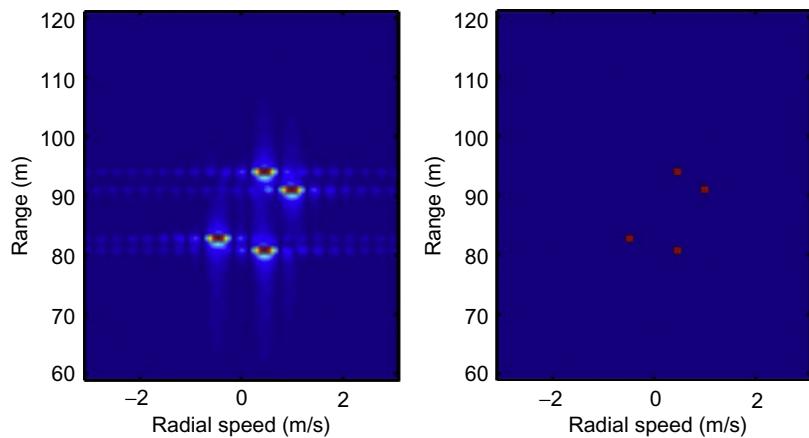
$$\min \|\mathbf{s}\|_1 \text{ subject to } \mathbf{Cs} = \mathbf{y}. \quad (6.25)$$

To confirm the improved performance of this discussed method let's assume that a standard pulse-Doppler radar transmits  $M$  chirped pulses with a bandwidth of  $B = 147$  MHz and a PRI of  $\tau_\square = 20$  ms. For the first case we consider that  $P$  is equal  $M = 20$  pulses and in the fast-time the Nyquist rate is fulfilled. Hence, all available data is used.

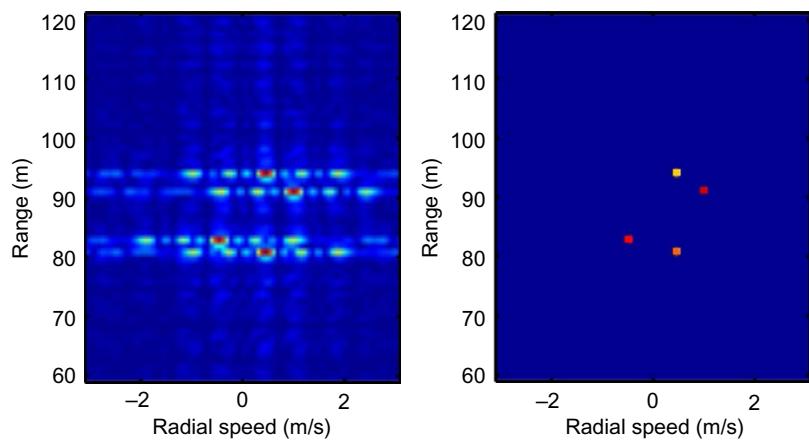
Fig. 6.2 shows the output of the simulation for the standard matched filter approach (left image) and the sparse matrix recovering method on the right.

For the co-prime sampling case  $P = 10$  pulses are randomly chosen from the  $M = 20$  transmitted pulses. Also the sampling rate in fast-time (range) is reduced by a factor of 2. The result is shown in Fig. 6.3.

The result obtained by the standard signal processing chain is shown in the left image. The direct comparison with the right image, which shows the solution obtained by the sparse reconstruction technique, confirms that the new approach is clearly able to estimate positions and velocity of all four moving targets with high accuracy. Furthermore, this method does not show any side-lobes.

**FIG. 6.2**

Result for a pulse-Doppler radar with a  $B = 147$  MHz,  $\tau_{\square} = 20$  ms, uniform sampling in fast- and slow-time with  $M = 20$  fulfilling Nyquist criteria. Left diagram shows the result of the matched filter and the right diagram obtained by sparse reconstruction technique for the case of four moving targets.

**FIG. 6.3**

Result for a pulse-Doppler radar with a  $B = 147$  MHz,  $\tau_{\square} = 20$  ms, nonuniform sampling in fast- and slow-time with  $P = 10, 11, \dots, 20$ . Left diagram shows the result of the matched filter and the right diagram of the compressive sensing case for four moving targets.

## 6.3 SPECTRAL SPARSITY

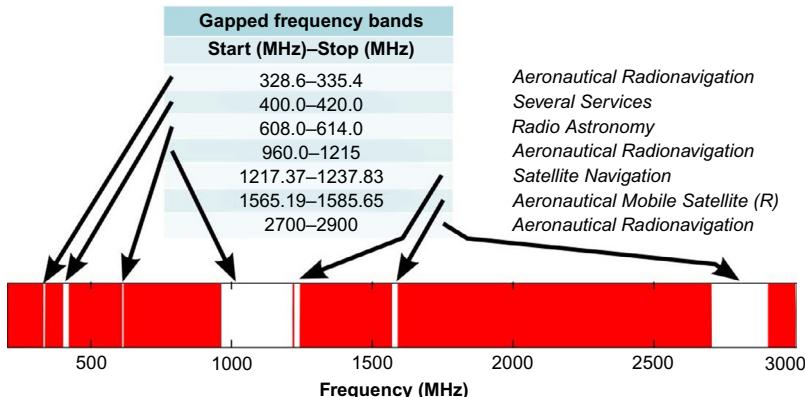
The sparse sampling scheme can also be applied to the frequency domain. The sparse reconstruction technique allows again to recover missing spectral information or to eliminate the influence of interference sources with a narrow frequency band.

### 6.3.1 RECOVERY OF MISSING OR CORRUPTED SPECTRAL INFORMATION

These days many applications use the electromagnetic spectrum for observation, navigation, and communication. For radar scientists who are interested in ultra-wideband radar systems, like ground penetrating radars (GPR) or high-resolution SAR systems, a large frequency bandwidth is needed. However, radio-regulation agencies have preallocated spectrum bands to various services, like GPS or mobile communication, and accordingly these frequency bands cannot be used for any radar purposes. Therefore a gapped frequency band, as shown in Fig. 6.4, have to be used for UWB radar, for instance.

These spectral gaps generate several sidelobes in the pulse compressed time-domain signal.

By adapting sparse recovering techniques it becomes possible to recover the missing information resulting from frequency gaps. Furthermore an adaptive approach allows to extract and to suppress highly fluctuating radio frequency



**FIG. 6.4**

Allocated frequency bands in the range up to 3 GHz.

Courtesy L.H. Nguyen, T.T. Do, T.D. Tran, Sparse model and sparse recovery with ultra-wideband SAR applications, in: 1st International Workshop on Compressed Sensing Applied to Radar (CoSeRa 2012), 2012, <http://workshops.fhr.fraunhofer.de/cosera>.

interference (RFI) signals without any prior knowledge about the frequency spectrum of the modulation scheme of these sources [32].

To recover missing spectral information let  $\mathbf{h}$  be the transfer function in the time-domain (in vector form) that corresponds to the gapped spectrum. The corrupted received signal is then the convolution between the undistorted signal and the transfer function of the gapped spectrum:  $\tilde{\mathbf{y}} = \mathbf{y} * \mathbf{h}$ . If the original received signal is specified by  $\mathbf{y} = \mathbf{A} \mathbf{s}$ , with  $\mathbf{A}$  the sensing matrix and  $\mathbf{s}$  the target state vector the radar signal with spectral gaps is

$$\tilde{\mathbf{y}} = (\mathbf{A} \mathbf{s}) * \mathbf{h}, \quad (6.26)$$

$$\tilde{\mathbf{y}} = (\mathbf{A} * \mathbf{h}) \mathbf{s}, \quad (6.27)$$

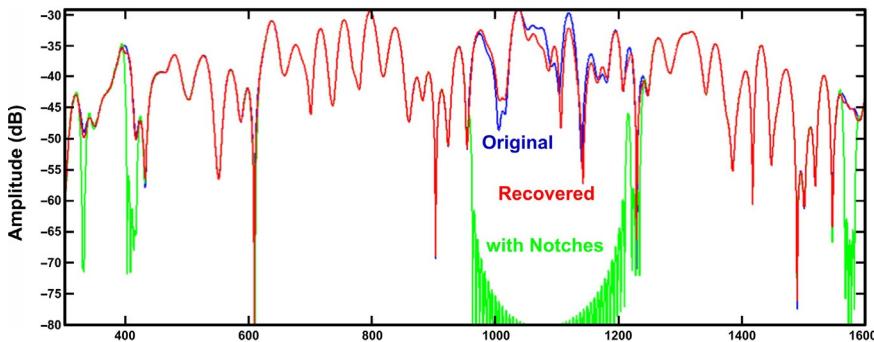
$$\tilde{\mathbf{y}} = \tilde{\mathbf{A}} \mathbf{s}, \quad (6.28)$$

where  $\tilde{\mathbf{A}} = \mathbf{A} * \mathbf{h}$  is the spectral gapped sensing matrix of the original matrix  $\mathbf{A}$  obtained by a column-by-column convolution with  $\mathbf{h}$ . Thus, the received signal (including the notched spectral content) is a linear combination of the delayed and weighted replicas of the notched versions of the time-shifted transmit waveforms. The coefficients  $\mathbf{s}$  can be calculated by solving the adapted minimization  $\ell_1$  problem:

$$\min_{\mathbf{s}} \|\mathbf{s}\|_1 \text{ subject to } \|\tilde{\mathbf{A}} \mathbf{s} - \tilde{\mathbf{y}}\|_2 < \sigma \quad (6.29)$$

with  $\sigma$  the threshold determined by the presence of noise.

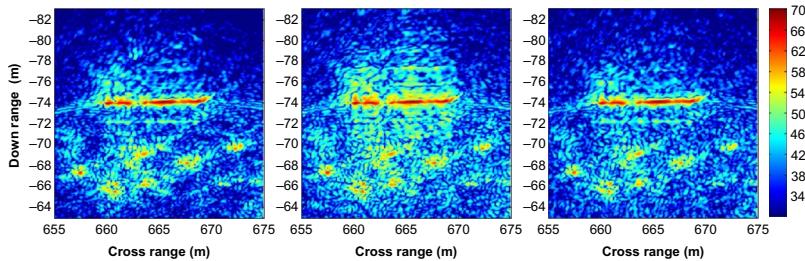
[Fig. 6.5](#) shows the result of a recovered spectrum which was corrupted by several notch filters.



**FIG. 6.5**

Recovered spectral information from a gapped frequency band.

Courtesy L.H. Nguyen, T.T. Do, T.D. Tran, *Sparse model and sparse recovery with ultra-wideband SAR applications, in: 1st International Workshop on Compressed Sensing Applied to Radar (CoSeRa 2012), 2012, <http://workshops.fhr.fraunhofer.de/cosera>.*

**FIG. 6.6**

Reconstructed SAR image using sparse reconstruction techniques. Left image shows the original SAR image, the middle image shows the corrupted and on the right side the recovered SAR image.

*Courtesy L.H. Nguyen, T.T. Do, T.D. Tran, Sparse model and sparse recovery with ultra-wideband SAR applications, in: 1st International Workshop on Compressed Sensing Applied to Radar (CoSeRa 2012), 2012, <http://workshops.fhr.fraunhofer.de/cosera>.*

This sparse recovering technique can also be applied to a 2D imaging radar, as demonstrated in Fig. 6.6.

The left image shows the SAR image obtained when there are no spectral gaps.

### 6.3.2 SUB- OR CO-PRIME SAMPLING IN THE SPECTRAL DOMAIN

This feature can be extended in such a way that a new radar type can be created. Instead of transmitting a chirped pulse with a bandwidth  $B$  the novel radar emits simultaneously only a few frequencies spread out over  $B$ , as proposed in [33]. This can be achieved by combining many different single frequency sources. Another approach to achieve a sparse populated frequency band is to use the orthogonal frequency-division multiplexing (OFDM) modulation scheme. The OFDM is well-established in the communication sector for wireless connections and it permits a high data throughput between two units. The OFDM signal consists of many different carrier frequencies with a typical separation of 1 kHz for applications at the telecommunication area. For a CS radar the distance between transmitted carriers can easily be expanded to several MHz.

Let a radar emit a signal consisting of  $M$  discrete  $f_1, \dots, f_M$  frequencies randomly distributed over a bandwidth  $B$ . The observed scene is sparsely populated by  $N$  targets. Clutter is not considered in this example. Each received signal can then be described by:

$$y_m = \sum_{n=1}^N s_n e^{-j 2\pi f_m \tau_n} \quad (6.30)$$

with  $s_n$  the complex amplitude taking the reflection coefficient of target  $n$  located at range  $r_n = \tau_n/2c_0$  and the propagation loss into account. Combining all measurements into a single vector  $\mathbf{y}$  yield:

$$\mathbf{y} = [y_1, \dots, y_M]^T = \mathbf{A}\mathbf{s} \quad (6.31)$$

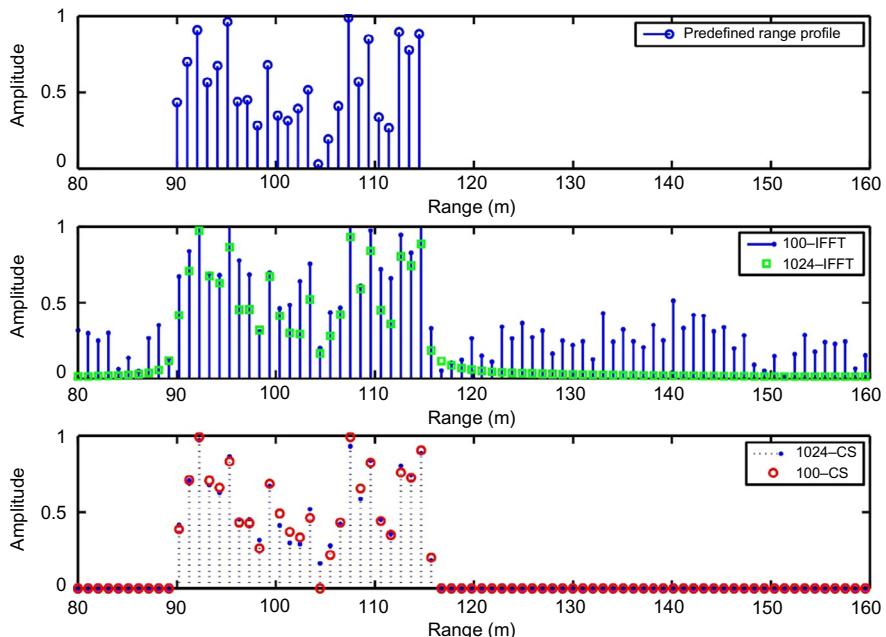
with the target state vector  $\mathbf{s} = [s_1, \dots, s_N]^T \in \mathbb{C}^{N \times 1}$  and the sensing matrix  $\mathbf{A} \in \mathbb{C}^{M \times N}$ :

$$\mathbf{A} = \begin{pmatrix} e^{j 2\pi f_1 \tau_1} & \dots & e^{j 2\pi f_1 \tau_N} \\ \vdots & \ddots & \vdots \\ e^{j 2\pi f_M \tau_1} & \dots & e^{j 2\pi f_M \tau_N} \end{pmatrix} \quad (6.32)$$

Solving the following  $\ell_1$  minimization optimization problem the sparse target state vector  $\mathbf{s}$  can be recovered:

$$\min_{\mathbf{s}} \|\mathbf{s}\|_1 \text{ subject to } \|\mathbf{A}\mathbf{s} - \mathbf{y}\|_2 < \sigma \quad (6.33)$$

**Fig. 6.7** shows the result obtained from such a radar emitting 1024 discrete frequencies uniformly distributed over a bandwidth of 147 MHz. The top diagram shows the predetermined range profile and the middle diagram the result obtained by the standard inverse Fourier transformation (IFFT). The circles (green) represents the result of the IFFT using all emitted frequencies and if we select randomly only 100 frequencies the situation changes dramatically, as shown by the stems (dots with blue line) in the second diagram.



**FIG. 6.7**

Comparison between traditional signal processing (inverse Fourier transformation, IFFT) and sparse reconstruction technique for different number of measurement samples.

The third lower graph shows the results obtained by sparse reconstruction techniques. Here the dots with dashed stems (blue) represent the solution when all emitted frequencies are considered. With only 100 frequencies the accuracy of the estimated range profile is reduced, however, the shape of the predefined range profile is clearly visible, as shown by the circles (red).

## 6.4 SPATIAL SPARSITY

A very interesting aspect of sparse reconstruction techniques can be seen in the possibility to reduce the number of sensor nodes without degrading the performance of the system. Sparsity in the spatial domain is accompanied by a significant reduction of the required hardware channels to collect the information from the surveillance area. The number of measurement nodes is directly related to the complexity of the system. Hence, spatial sparsity combined with sparse reconstruction techniques can reduce the price dramatically.

A prominent example for such an application is the DOA estimation with a sparse antenna array. If we know the emitted signal structure of the point sources, which can be an active radar or a target echo illuminated by such a device, we can estimate the DOA of the received signals. However, with two elements we can only estimate the angle if there exists one single source. Increasingly the number of receive elements will increase the estimation accuracy of the direction and the possibility to distinguish between several sources.

If several signal sources are present sparse reconstruction techniques will be an alternative to the classical spatial superresolution methods as periodogram, correlogram, Generalized Cross Correlation (GCC), and Capon. Parametric approaches are Minimum Variance Distortionless Response (MVDR), MUSIC (MULTiple SIGnal Classification), and ESPRIT (Estimation of Signal Parameters via Rotational Invariance Technique) [34].

With sparse reconstruction techniques we can overcome some of the limitations of sparse array processing caused by nonoptimal number of elements and their spacing. For instance, the ability to resolve closely spaced sources is improved and the sensitivity against correlated signal sources is increased. One of the advantages of this technique is that it allows to reduce the spatial sampling without declining the DOA estimation accuracy.

### 6.4.1 DIRECTION-OF-ARRIVAL (DOA)

Let a known waveform  $s(t)$  impinge an antenna array consisting of  $M$  receive elements. All related target parameters like range and Doppler have already been estimated during the first appropriate signal processing stages before DOA estimation is performed. The positions of the receive nodes can have an arbitrary geometry and are described in general by  $\xi_m = [x_m, y_m, z_m]^T$ , as depicted in Fig. 6.8.

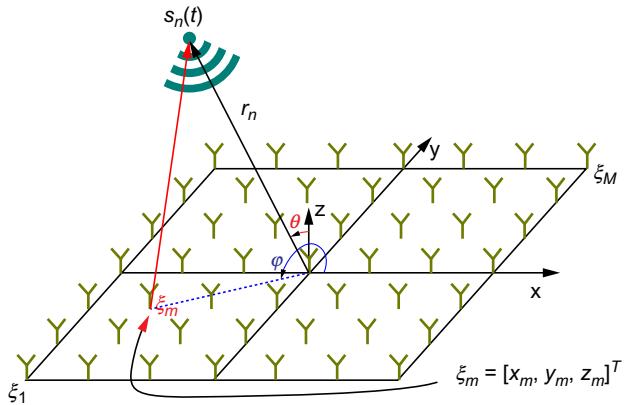


FIG. 6.8

Antenna array.

The search area of a general antenna array is described by the elevation angle  $\varphi_n$  and the azimuth angle  $\theta_n$  as the signal sources are in the far-field. In a first approximation the received signal at each antenna element can be modeled by considering an additional phase shift caused by the extra traveling path compared to the reference element. Then we can write for the  $N$  impinging signals at receive node  $m$ :

$$y_m(t) = \sum_{n=1}^N s_n e^{-j k_r \|r_n - \xi_m\|_2} \quad (6.34)$$

with  $k_r = 2\pi/\lambda$  the wave number. For the final array we combine all measurements into a single vector  $\mathbf{y}$  and if  $\mathbf{s}(t) = [s_1, \dots, s_N]^T$  denotes the received signal strength from all search angles we get:

$$\mathbf{y}(t) = [y_1(t), \dots, y_M(t)]^T = \mathbf{A} \mathbf{s}(t). \quad (6.35)$$

The sensing matrix  $\mathbf{A}$  combines the transfer functions of all receive nodes:

$$\mathbf{A} = \begin{pmatrix} e^{-j k_r \|r_1 - \xi_1\|_2} & \dots & e^{-j k_r \|r_N - \xi_1\|_2} \\ \vdots & \ddots & \vdots \\ e^{-j k_r \|r_1 - \xi_M\|_2} & \dots & e^{-j k_r \|r_N - \xi_M\|_2} \end{pmatrix}. \quad (6.36)$$

As we know that the number of signal sources  $K$  is less than the number of receive elements  $M$  the underdetermined linear system described by Eq. (6.35), with  $M < N$ , can be solved by applying the sparse reconstruction technique.

$$\min_{\mathbf{s}} \|\mathbf{s}\|_1 \text{ subject to } \|\mathbf{A} \mathbf{s} - \mathbf{y}\|_2 \leq \sigma. \quad (6.37)$$

Robustness and accuracy of the estimation can be enhanced if multiple time samples are combined into a single measurement, as proposed in [35].

#### 6.4.1.1 DOA with a linear array

For the first example we assume five narrowband sources located in the far field at the angles  $30^\circ$ ,  $45^\circ$ ,  $50^\circ$ ,  $100^\circ$ , and  $130^\circ$ , respectively. The plane waves impinge on an uniformly linear antenna array of  $M = 30$  elements equally spaced by a half wavelength  $d = \lambda/2$ . Each receive node acquires  $L = 200$  samples to enhance the signal-to-noise ratio (SNR) by 10 dB.

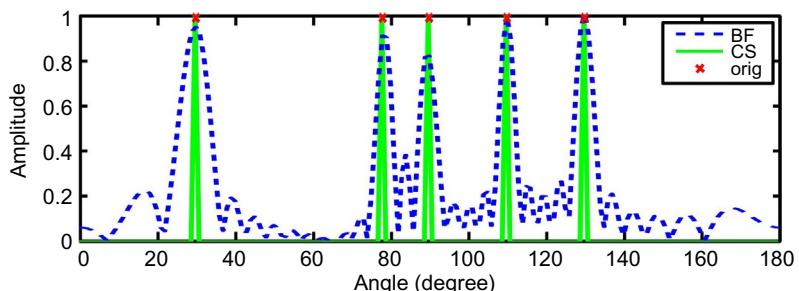
[Fig. 6.9](#) shows the result from this simulation. The signal sources are marked by red crosses (dark gray crosses in print versions). The output of the traditional beamformer approach is shown by the blue dotted line (dark gray dotted line in print versions).

The green line (gray line in print versions) represents the estimation obtained by sparse reconstruction. Particularly at closely spaced sources the advantage of the new technique is obvious.

#### 6.4.1.2 DOA with a 2D array

The advantage of this new technique is apparent when an antenna array is sparse populated with randomly selected positions of the antenna elements. Consider a two-dimensional circular antenna array, as depicted in the left diagram in [Fig. 6.10](#). The blue crosses denote the element positions of a full populated array with  $N_{full} = 648$  elements. Compared to this huge amount of elements we select randomly only 80 active antennas for this simulation. These active elements are marked by circles (in the left top diagram). Also shown are the positions of possible targets by circles in the right top diagram.

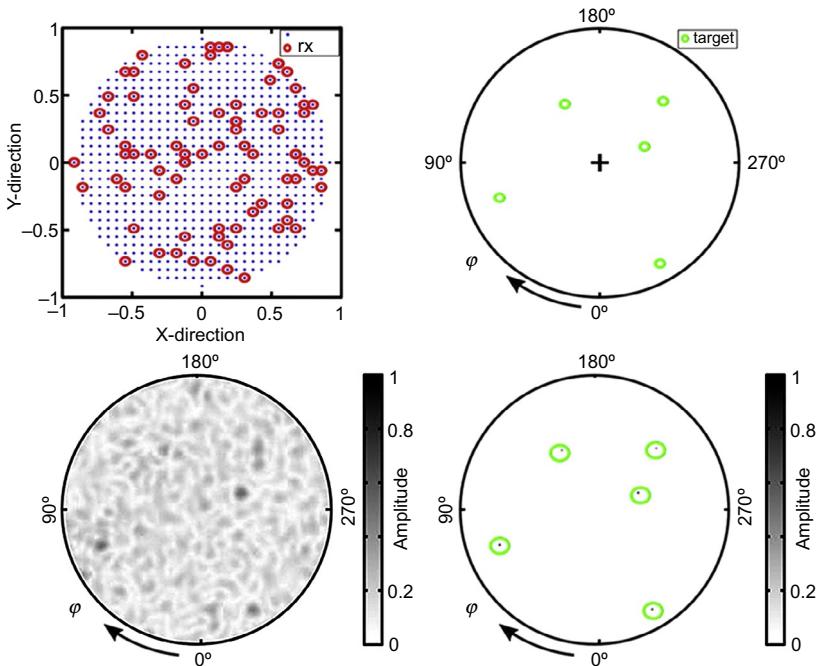
The SNR is 20 dB and the received signals from all five targets have the same amplitude for this investigation. The top left image shows the antenna configuration and the right image on the same row the positions of the five targets projected onto the x/y-plane. The lower left image shows the determined amplitude distribution over the azimuth/elevation plane obtained by the classical beamformer approach and the lower right image by the sparse reconstruction technique.



**FIG. 6.9**

---

Linear antenna array with  $N = 30$  elements equally spaced by  $d = \lambda/2$  and 5 signal sources.

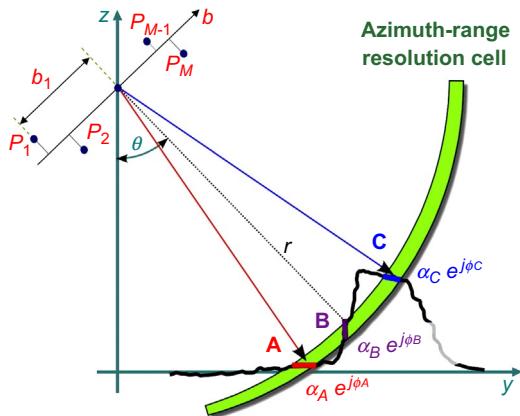
**FIG. 6.10**

Circular antenna array with  $N = 80$  randomly selected elements and five targets maneuvering in the surveillance space. The top left image clockwise shows the x/y-plane and the antenna grid with the selected receive nodes. The right image shows the five targets in the hemisphere projected onto the x/y-plane. The result obtained by the beamformer approach is depicted in the lower left image and the right image shows the result obtained by sparse reconstruction techniques.

### 6.4.2 3D-SAR

An important application of this new reconstruction technique can be seen in repeat-pass differential interferometry SAR (DInSAR) [36] and tomographic SAR, also referred to 3D SAR imaging.

These methods are based on the principle that each SAR acquisition, which corresponds to different measurement paths, can be associated with another element of a linear antenna array parallel to the elevation direction. After applying classical synthetic aperture techniques for each acquisition track a virtual antenna array in elevation direction can be synthesized. Fig. 6.11 sketches such a situation where different antenna positions are labeled by  $P_1, \dots, P_M$ . Due to this antenna array in elevation direction a height profile can be determined from the scatterer distribution in the elevation direction and, hence, a full 3D image of the scene can be generated [37, 38]. As the measurements were performed at different times, this

**FIG. 6.11**

Possible scatterer distribution in a range-azimuth pixel of a single-look SAR image.

technique also allows to measure the movement of each scatterer (velocity). Such a technique referred to as differential SAR tomography is also well-known as 4D (space-velocity) SAR imaging [39, 40].

It would be desirable if tomographic SAR had an isotropic spatial resolution in range, azimuth, and elevation. From this it follows that for space-based SAR systems the synthesized antenna array in elevation must have an aperture of several kilometers. However, modern high-resolution SAR satellites use a very tight orbit control which only allows an interferometric aperture of about 300 m or less, which implies a resolution in elevation  $\Delta r_h$  of about 10–50 times less than in range  $\Delta r_r$  or azimuth  $\Delta r_x$  [41].

A realistic assumption regarding the scene is that each azimuth-range bin contains only a few scatterers. Thus we can apply sparse reconstruction techniques with its high resolution capabilities for the subsampled aperture to recover these few scatterers [42].

The first processing stages of an interferometric SAR processing chain are carefully focusing the acquired SAR images in azimuth and range, undertake an image registration of all data to a reference master image, compensate geometric phase caused by atmospheric disturbances related to a reference point. After these steps a generic azimuth-range pixel acquired at the  $m$ th measurement is denoted by  $x_m$ , following the 3D notation from Fornaro et al. [38]. Each azimuth-range pixel is a superposition of a few scatterers distributed in the elevation direction  $h$  which can be described by the complex amplitude distribution  $s(h)$ . Also a deformation term has to be considered whose line-in-sight component is equal to  $d(h, t_m)$  [38]:

$$x_m = \int_{\Delta s} s(h) e^{-j 2 \pi \xi_m h} e^{-j \frac{4 \pi}{\lambda} d(h, t_m)} dh. \quad (6.38)$$

The function  $s(h)$  models the complex amplitude of the backscattering in the selected azimuth-range pixel. The sampling spatial frequency  $\xi_m$  [1/m] in the elevation direction is given by:

$$\xi_m = \frac{2 b_m}{\lambda r} \quad (6.39)$$

with  $b_m$  the orthogonal baseline component of the  $m$ th acquisition pass (see Fig. 6.11),  $\lambda$  the center wavelength of the transmitted waveform, and  $r$  the range to the scene point.

By expanding the exponential deformation term in Fourier harmonics and using the relation  $d(h, t_m) = \eta_m v$ , Eq. (6.38) can be rewritten as:

$$x_m = \int_{\Delta h} \int_{\Delta v} s(h, v) e^{-j 2 \pi (\xi_m h + \eta_m v)} dh dv, \quad (6.40)$$

where  $s(h, v)$  is now the backscattering distribution projection in the elevation/velocity plane and  $\eta_m$  [s/m] the temporal frequency:

$$\eta_m = \frac{2 t_m}{\lambda}. \quad (6.41)$$

In general the measurements are nonuniformly sampled in space and time of the backscattering distribution in the elevation/velocity plane.

Discretizing Eq. (6.40) by  $N$  points corresponding to discrete bins ( $s(h_i, v_j)$ ) from the elevation/velocity-plane we obtain the following matrix problem:

$$\mathbf{x} = \mathbf{A} \mathbf{s} + \boldsymbol{\epsilon} \quad (6.42)$$

with:

- $\mathbf{x} = [x_1, \dots, x_M]^T$  the vector collecting the  $M$  acquired measurements of the selected discrete azimuth-range bin and  $T$  is the transposition vector.
- $\mathbf{s} = [s_1, \dots, s_N]^T$  the vector which represents values of the backscattering distribution at the discrete bins ( $h_i, v_j$ ).
- $\mathbf{A} = [\mathbf{a}_1, \dots, \mathbf{a}_N]$  the  $M \times N$  sensing matrix which collects the  $N$  steering vectors given by:  $\mathbf{a}_n = \mathbf{a}(h_i, v_j) = \mathbf{a}_n = \exp\{j 2 \pi (\xi h_i + \eta v_j)\} / \sqrt{M}$  with:  $\xi = [\xi_1, \dots, \xi_N]^T$  being the vector collecting the spatial frequencies,  $\eta = [\eta_1, \dots, \eta_N]^T$  the vector collecting the temporal frequencies.
- $\boldsymbol{\epsilon}$  the  $M$ -dimensional noise vector. The noise is modeled as independent circular complex Gaussian process with zero mean and covariance matrix  $\sigma_m^2 \mathbf{I}_M$ , where  $\sigma_m$  denotes the noise power, i.e.,  $\boldsymbol{\epsilon} \sim \mathcal{C}\mathcal{M}(0, \sigma_m^2 \mathbf{I}_M)$ .

#### 6.4.2.1 Experimental results

To verify the proposed sparse signal reconstruction scheme it has been applied to real data consisting of 25 spotlight Terra-SAR-X images.

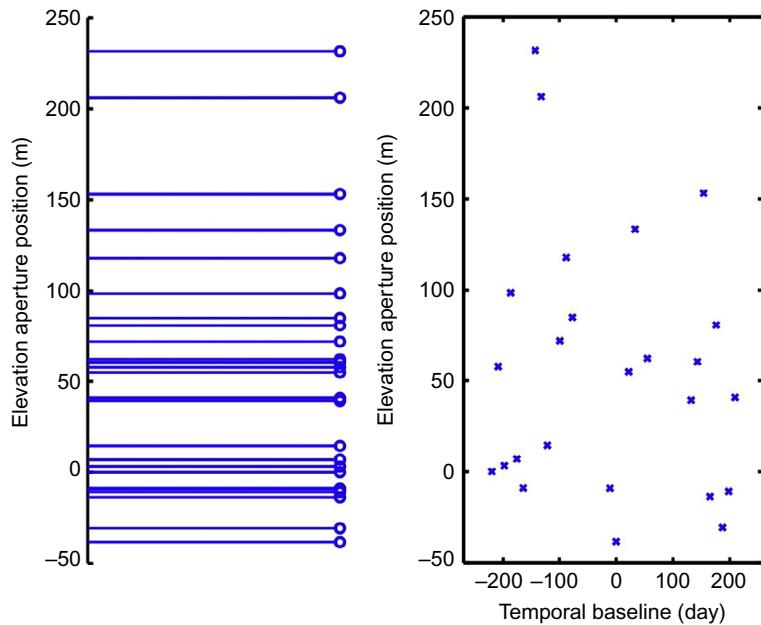
The data were recorded over the area from Las Vegas, NV, USA with a nominal revisit period of 11 days. The resolution in slant-range was 0.6 m, which corresponds to a ground range resolution of about 1 m at an incidence angle of  $35.8^\circ$ , and in

**FIG. 6.12**

On the left an optical image (Google Earth) and on the right a TerraSAR-X single look SAR image from the Mirage hotel, Las Vegas, NV, USA.

azimuth 1 m. Fig. 6.12 shows on the left an optical image acquired by Google Inc. and a single-look SAR image obtained by TerraSAR-X on the right side.

The satellite flight path was controlled in such a way that the diameter of the orbital tube was less than 500 m. Fig. 6.13 shows the resulting elevation aperture in space and time. The synthetic antenna aperture in elevation is about  $\Delta b = 269.9$  m.

**FIG. 6.13**

Elevation aperture in time and space.

In the case of an virtual aperture fulfilling the Shannon-Nyquist criteria the achievable elevation resolution is:

$$\Delta r_e = \frac{\lambda r_0}{2 \Delta b} = 35.6 \text{ m.} \quad (6.43)$$

At an incidence angle of  $35.8^\circ$  this corresponds to a height resolution (z-direction) of about 20.9 m.

The standard beamformer approach uses the conjugate operator of  $\mathbf{a}_n$  to reconstruct the unknown function  $\boldsymbol{\alpha}$ :

$$s(h_i, v_j) = \mathbf{a}_n^H \mathbf{x} \quad (6.44)$$

with  $\mathbf{a}^H$  the Hermitian matrix operator. We have to choose between the two hypotheses without and with a scatterer

$$\begin{aligned} \mathcal{H}_0 : \mathbf{x} &= \boldsymbol{\epsilon}, \\ \mathcal{H}_1 : \mathbf{x} &= \mathbf{a}_n \boldsymbol{\alpha} + \boldsymbol{\epsilon}, \end{aligned}$$

where  $\mathbf{a}_n$  is the steering vector. Determining for all considered height and velocity combination the general likelihood ratio test (GLRT)

$$\begin{aligned} (\hat{h}, \hat{v}) &= \arg \max_{h, v} \| \mathbf{a}^H(h, v) \mathbf{x} \|_2^2 \\ &= \arg \max_{h, v} \frac{ \| \mathbf{a}^H(h, v) \mathbf{x} \| }{ \| \mathbf{a}(h, v) \|_2 \cdot \| \mathbf{x} \|_2 } \stackrel{\mathcal{H}_1}{\stackrel{\mathcal{H}_0}{\lessgtr}} \mathcal{T} \end{aligned}$$

with  $\mathcal{T}$  the detection threshold. The bin  $(i, j)$  with the maximum value yields the useful measurement for target localization and velocity  $(\hat{h}_i, \hat{v}_j)$ .

To improve the estimation process, in particular for the sparse reconstruction technique, only stable scatterers will be considered. Due to the time interval of 11 days between each successive measurement scatterer in an azimuth-range bin might emerge. As a result from this one has to estimate for each elevation-velocity combination  $(h_i, v_j)$  the coherent measurements which will then be considered to estimate the scatterer parameters. Firstly the correlation matrix  $\mathcal{C}_\varphi$  for each elevation-velocity bin  $(h_i, v_j)$  is determined:

$$\mathcal{C}_\varphi = \frac{1}{N} (\mathbf{a}_n^H \circ \mathbf{y}) (\mathbf{a}_n^H \circ \mathbf{y})^H \quad (6.45)$$

with  $\circ$  the Hadamard (elementwise) product of two vectors. For each element of the correlation matrix the following test is carried out:

$$|\Delta(\mathcal{C}_\varphi(i, j))| = \begin{cases} 1 : & < \mathcal{T}_\varphi \\ 0 : & \text{else} \end{cases}$$

with  $T_\varphi$  the maximum allowable phase deviation of a stable scatterer. The sum over each row yields the maximum number of coherent measurements  $N_{coh}$  for this elevation/velocity-bin:

$$N_{coh} = \arg \max \left( \sum_{i=1}^N |\varphi(\mathcal{C}_\varphi(i, j))| \right).$$

If  $N_{coh}$  is greater than a specific threshold  $T_{coh}$  we can assume that a persistent scatterer exists. The final estimation of the scatterer parameters is then carried out exclusively with these coherent measurements determined by  $N_{coh}$ .

The left image in Fig. 6.14 shows the maximum  $N_{coh}$  obtained by the described procedure for each azimuth-range bin for a phase deviation of  $45^\circ$ . The right image shows the azimuth-range plane for a  $T_c \geq 17$  indicating where at least one stable scatterer exist.

To estimate the scatterer parameters within the sparse reconstruction technique the following linear minimization problem has to be solved:

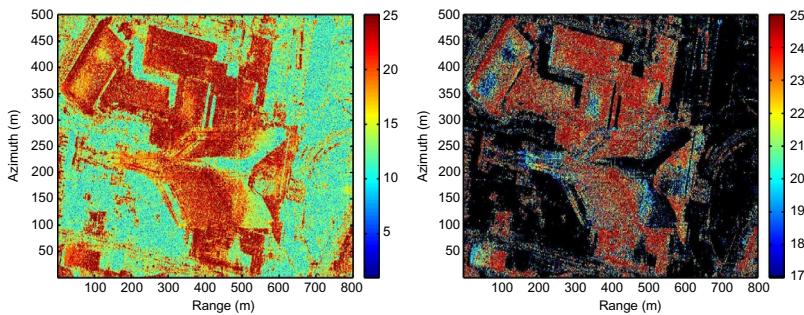
$$\min_s \|s\|_1 \text{ subject to } \|As - x\|_2 \leq \sigma \quad (6.46)$$

with  $\sigma$  the threshold defined by the presence of noise provides the point scatterer distribution in the elevation/velocity-plane.

Fig. 6.15 shows for comparison the estimated height using the traditional matched filter approach (left image) and the sparse reconstruction technique (right image) where only the top scatterers are shown.

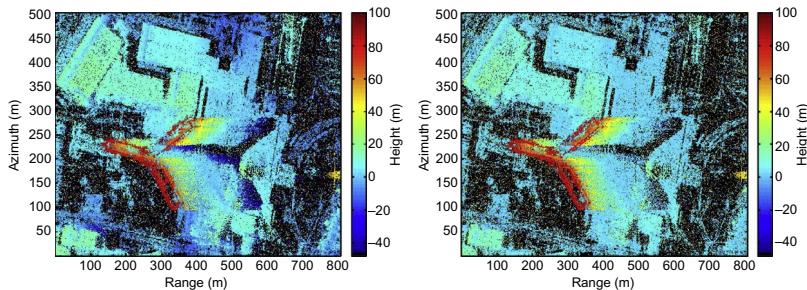
Focusing on the region of the façade of the Mirage hotel it is evident that the sparse reconstruction can distinguish between scatterers from the furrowed façade and from the ground.

Fig. 6.16 shows the number of scatterers contributing to each azimuth-range cell estimated by CS.

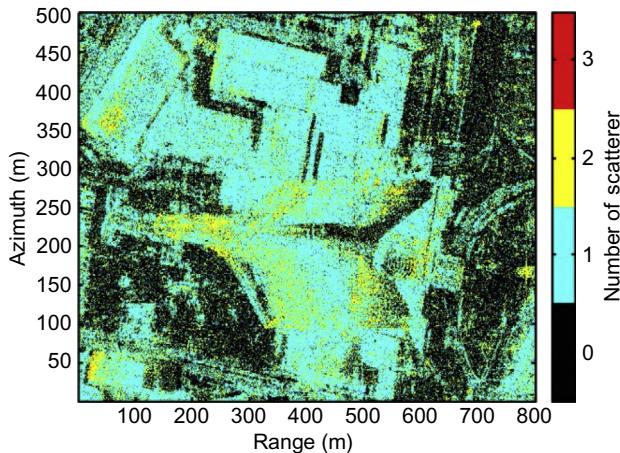


**FIG. 6.14**

Max number of coherent measurements on the left side with a maximum of phase deviation of  $45^\circ$ . The right shows the same result, however, the number of coherent measurements is limited to a minimum of 17 (the phase deviation is still  $45^\circ$ ).

**FIG. 6.15**

Estimated topography by the nonparametric MF method on the left image and by the sparse reconstruction technique on the right image. For the CS estimation only the top scatterer is shown.

**FIG. 6.16**

Estimated number of scatterers resolved by CS.

## 6.5 GROUP SPARSITY

Another very interesting feature of the sparse reconstruction technique is the ability to form groups containing information from different distributed sensors observing the same surveillance area and to estimate target parameters from all these sensors. Let's consider a MIMO sensor system consisting of several distributed transmit and receive nodes. For a MIMO system with homogeneous sensors, as it is the case for multistatic radar/sonar systems, we can form from the distributed network many mono-/bistatic constellations. It has been shown that detection and tracking performance of a MIMO system can be strongly improved by combining the information gained by all transmit-receive nodes [43–45].

Let us assume that the surveillance area is sparsely populated by  $K$  objects and observed by  $T \cdot R$  transmit-receive pairs. According to this target  $k$  can then be described by a  $L$ -tuple of complex amplitudes  $x_k^{(l)}, l \in [L = T \cdot R]$  where the superscript  $(l)$  denotes the index of the related transmit-receive combination,  $T$  the number of transmitters, and  $R$  the number of receivers, respectively. Each target state point  $x_k^{(l)}$  of this tuple describes the object by the same coordinates, velocities, directions, and others parameters [46].

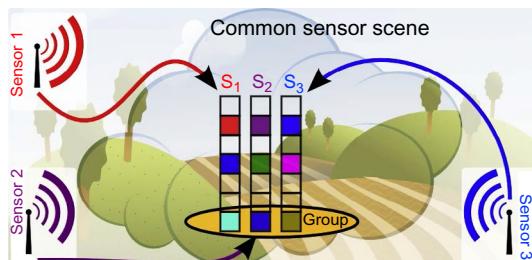
As each tx-rx pair observes the same target from different aspects angles the corresponding tuple consists of different complex values due to the dependency of the RCS on the aspect angle and the different range attenuation. However, if a target exists all the corresponding target state point tuples have nonzeros entries, as depicted in Fig. 6.17.

In the following we restrict us to a finite number of target state points  $p \in [P]$ . Then we can describe the observed scene by:

$$\mathbf{S} = \begin{pmatrix} s_1^{(1)} & \dots & s_1^{(L)} \\ \vdots & \ddots & \vdots \\ s_P^{(1)} & \dots & s_P^{(L)} \end{pmatrix}, \quad (6.47)$$

where the  $l$ th column vector  $\mathbf{s}^{(l)} = [s_1^{(l)}, \dots, s_P^{(l)}]^T$  contains the observed information of sensor  $l$  about the received echo amplitudes from target with the parameter states  $p = 1, \dots, P$ . In contrast the  $p$ th row vector  $\mathbf{x}[p] = [x_p^{(1)}, \dots, x_p^{(L)}]$  contains the amplitudes for a specific target state point  $p$  observed by all  $L$  sensors.

This allows us to form for each row a group of identical target states containing only the corresponding measurements from all contributing sensors. If a target exists all members of the corresponding group show an entry in contrast to groups without targets. This is depicted in Fig. 6.17 by the circle.



**FIG. 6.17**

If a target exists all transmit-receive pairs will show a measurement correlated to its state. This principle of the group sparsity is depicted by the circle combining the measurement vectors of all nodes.

### 6.5.1 GROUP MODEL

A scene is illuminated by  $T$  transmitters, for simplicity we restrict us to a two-dimensional space, located at  $\mathbf{s}_{tx_t} = [x_{tx_t}, y_{tx_t}]^T$ , and the scene is observed by  $R$  receivers at the locations  $\mathbf{x}_{rx_r} = [x_{rx_r}, y_{rx_r}]^T$ . This corresponds to  $L = T \cdot R$  different measurements for each scene point:

$$\mathbf{y}^{(l)} = \mathbf{A}^{(l)} \mathbf{s}^{(l)} + \mathbf{n}^{(l)}. \quad (6.48)$$

The sensing matrix  $\mathbf{A}^{(l)} \in \mathbb{C}^{M^{(l)} \times P}$  is related to the  $l$ th transmit-receive combination,  $M^{(l)}$  is the number of measurements taken by this sensor pair,  $P$  the number of considered target state points, and  $\mathbf{n}^{(l)}$  the noise vector.

Combining the measurements from all  $L$  sensor pairs into a single measurement vector yields:

$$\tilde{\mathbf{y}} = \mathbf{A}^{\ddagger} \tilde{\mathbf{s}} + \tilde{\mathbf{n}} \quad (6.49)$$

with  $\tilde{\mathbf{y}} = \text{vec}(\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(L)})$ ,  $\tilde{\mathbf{n}} = \text{vec}(\mathbf{n}^{(1)}, \dots, \mathbf{n}^{(L)})$ , and  $\tilde{\mathbf{s}} = \text{vec}(\mathbf{s}^{(1)}, \dots, \mathbf{s}^{(L)})$  the staggered vectors relating to the sensors and  $\mathbf{A}^{\ddagger}$  the block diagonal matrix constructed by the sensing matrices  $\mathbf{A}^{(L)}$ :

$$\mathbf{A}^{\ddagger} = \begin{pmatrix} \mathbf{A}^{(1)} & 0 & 0 \\ 0 & \mathbf{A}^{(l)} & 0 \\ 0 & 0 & \mathbf{A}^{(L)} \end{pmatrix}. \quad (6.50)$$

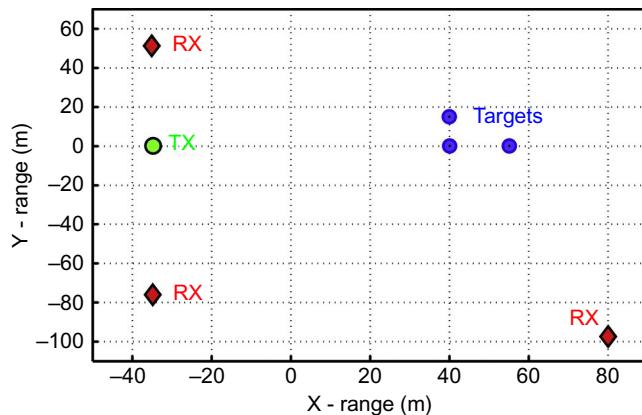
Several promising sparse reconstruction algorithms exist which are able to cope with group sparsity, for instance group lasso [47], block orthogonal matching pursuit (BOMP) [48], group-sparse BPDN, or even the SPGL1 offer the possibility to form groups [14]. Let  $\mathbf{s}_{\alpha_p}$  denote the target state tuple  $p$  consisting of  $\mathbf{s}_{\alpha_p} = [s_p^{(1)}, \dots, s_p^{(L)}]$  we have to solve the following minimization problem:

$$\min_i \sum_i \|\tilde{\mathbf{s}}_{\alpha_i}\|_1 \text{ subject to } \|\mathbf{A}^{\ddagger} \tilde{\mathbf{s}} - \tilde{\mathbf{y}}\|_2 < \sigma \quad (6.51)$$

with the threshold  $\sigma$  restricted by noise.

### 6.5.2 EXAMPLE: SIMO RADAR NETWORK

A scene is illuminated by one transmitter and observed by three spatial distributed receivers, as depicted in Fig. 6.18. For this simulation we assume that the transmitter is a WLAN access point and we consider only the beacon signal which is emitted regularly. The frequency bandwidth of the beacon signal is 20 MHz which corresponds to a monostatic range resolution of 7.5 m and the modulation scheme is an OFDM. An addition Barker code modulation ensures that the transmitted signal energy is spread over the whole frequency bandwidth even with no information change in the data stream.



**FIG. 6.18**

Simulation setup consisting of one TX, three RX, and three targets.

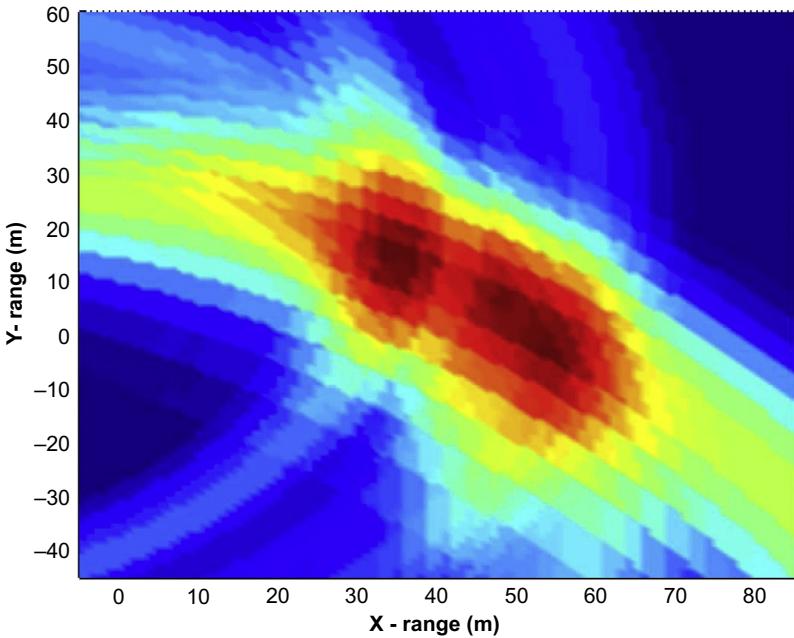
The scene consists of three point targets with identical radar cross section, separated in such a manner that the minimum range difference of the received signals is greater than 15 m. As all targets are separated more than twice the range resolution of the radar system and they should clearly be resolved after the signal processing.

The sample rate was set to 60 MHz which corresponds to an oversampling factor of 3. The duration of a beacon frame is 1.4 ms and the number of samples per frame is 84,000.

The first step in the matched filter approach, described in detail in [49], is to determine a digital replica of the original transmitted signal from the acquired wifi beacon frame data. To reduce the effect of the Barker modulation, which takes place during the coding process, a Barker-sidelobe-reduction (BSLR) filter is applied in the subsequent processing stage [50]. After pulse compression and BSLR filter the output of each transmit-receive-pair is transformed from the individual range-Doppler plane to a common Cartesian coordinate system for data fusion. Fig. 6.19 shows the result obtained by an incoherent integration. The length of the beacon frame was  $L = 84,000$  samples.

It is evident from this figure that the MF approach is not able to resolve all three targets. The reason is that the impulse responses from two objects merge in such a way that they cannot be separated any more.

For the sparse reconstruction we firstly determine the sensing matrices for all transmit-receive combinations  $A^{(l)}$  and arrange them as described by Eq. (6.50) into a single sensing matrix. The size of the sensing matrix depends on the realization of the pulse compression, if the compression is performed before the sparse reconstruction technique takes place or if the compression is realized within the sparse reconstruction scheme. To show the advantage of the new technique the sensing matrix is realized in this considered multistatic simulation with noncompressed data, as proposed in Section 6.2.

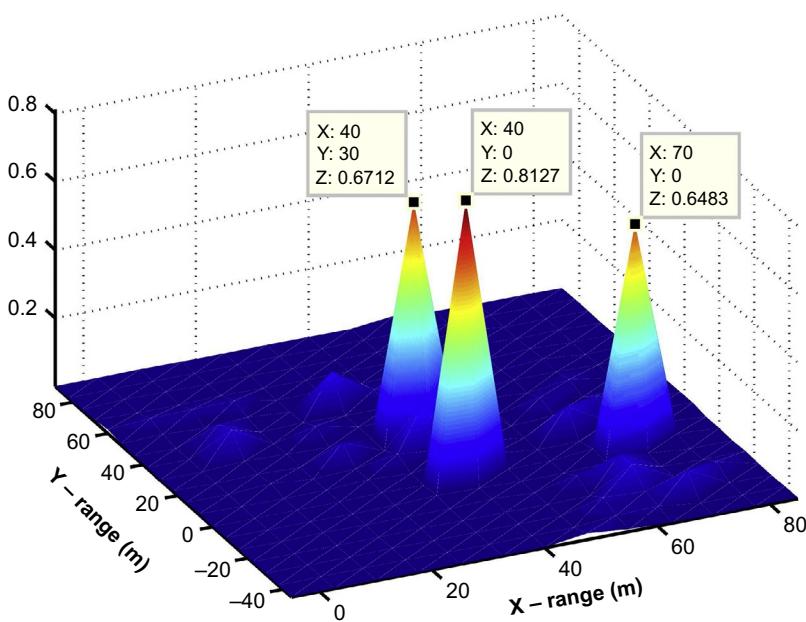
**FIG. 6.19**

Target positions determined by the traditional MF approach and  $L = 84,000$  measurements from each Tx-Rx-pair.

After setting up the corresponding groups it is possible to solve optimization task described by Eq. (6.51). As the pulse compression is also implemented into the sensing matrix the sparse reconstruction technique is able to estimate the correct number of targets and their positions even with less samples. The result is shown in Fig. 6.20 where only 800 samples, randomly selected from the whole frame dataset of 84,000 samples. This corresponds to a reduction in the data rate of a factor of 105.

Compared to the MF approach the group sparsity technique is able to recover all three targets in the scene.

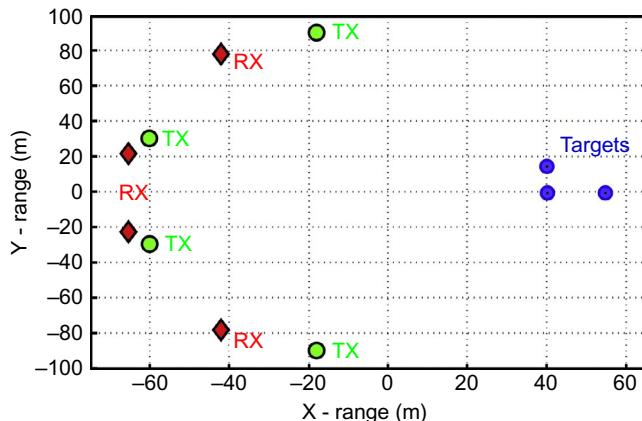
The reason for the superior performance in spite of the high data rate reduction is the digital modulation of the communication/beacon signal (OFDM with an additional Barker code modulation) which spreads the spectral energy of each symbol over the entirely channel spectrum of 20 MHz. Under the constraint that the SNR is high enough this technique enable an efficient implementation of the sparse reconstruction technique on uncompressed signals. Furthermore there is no need to implement a BSLR filter as the effect of the Barker modulation is already incorporated in the sensing matrix  $\mathbf{A}$ . Each columns  $p$  of the sensing matrix  $\mathbf{A}^{(l)}$  is a time delayed replica of the reference signal with a corresponding time delay/Doppler shift of target state  $s_p$ . Thus the sparse reconstruction method estimates the scene by taking all these modulation effects into account.

**FIG. 6.20**

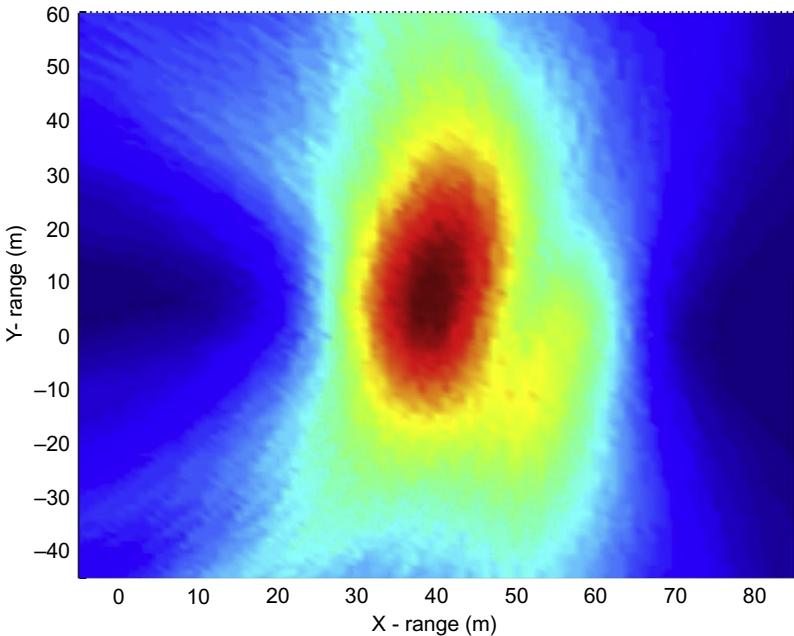
Reconstructed state vector using group sparsity of SGPL1 and with  $L = 800$  from each TX-RX-pair.

### 6.5.3 EXAMPLE: MIMO RADAR NETWORK

For the second example we assume a MIMO radar network consisting of four transmitters and four receivers located on a semicircle. Three targets in a group are separated by 15 m, as depicted in Fig. 6.21.

**FIG. 6.21**

Simulation setup consisting of four TX, four RX, and three targets.

**FIG. 6.22**

Target positions determined by the matched filter approach.

As in the SIMO case the target aren't all detected by the matched filter method, as shown in Fig. 6.22.

By the use of sparse reconstruction techniques all three targets were detected and their positions estimated with high accuracy. As the number of transmit-receive-pairs is increased the number of samples from each sensor pair can be further reduced. Fig. 6.23 shows the result obtained with  $L = 50$  samples per TX/RX-pair.

#### 6.5.4 EXAMPLE: SFN RADAR

For the last example we assume a single frequency network (SFN) consisting of three DVB-T stations and a single receiver. All transmitters coherently emit the same information using the identical frequency channel. This is a very attractive constellation due to its low costs of operation, maintenance and a compact size of the receive node. As all signals are transmitted using the same frequency band the receiver can be designed to use a single receive frequency bandwidth. This advantage, however, has the drawback that one target is illuminated by different transmitters and therefore generate several echoes. These echoes are interlaced in the received signal due to the different ranges from the various transmitters to the target. In the case of a single

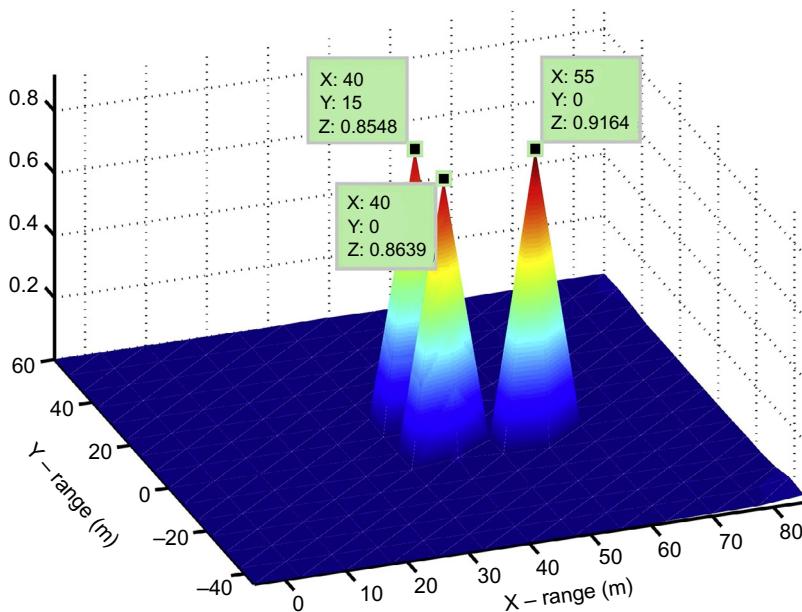


FIG. 6.23

Target state vector obtained by SGPL1 and  $L = 50$  samples from each TX/RX-pair.

target the ambiguity can be resolved without problems. However, if there are several objects then traditional methods will bring up many ghost targets, which requires to employ sophisticated tracking techniques.

Applying the groups' sparse reconstruction technique the problem of interlaced echoes can be resolved and the target states can be accurately estimated.

In real scenarios targets do not act like point scatterers. In general they are composed of multiple scattering centers distributed over the target surface. To improve the detection rate we integrate coherently over a large number of OFDM symbols [51]. Furthermore, to obtain a high Doppler resolution a Fourier transformation over 1023 pulses has to be conducted in slow-time. As a consequence of these steps the received signal shows an outstretch of a moving target echo over several range/Doppler cells. For high-resolution radars this is even more valid. To incorporate this effect in the target detection and parameter estimation process a cluster database is set-up which contains a link between detections and their related range/Doppler bins.

To achieve the desired clustering it is assumed throughout the CFAR detection process that echoes from two different targets do not commingle in the range/Doppler map. All connected *range/Doppler-bins* of an extended target are described by the vector  $\tilde{\mathbf{y}}$ .

In the fusion and estimation stage an adapted group sparsity algorithm allows us to take extended targets into account. In addition to the group sparsity, which describes the relation between target position and its interlaced echo pattern, a cluster database, which comprises the vectors  $\tilde{\mathbf{y}}_l$ ,  $l \in 1, \dots, L$ , where  $L$  is the number of detections, is introduced.

$$\mathbf{y} = \sum_{l=1}^L \tilde{\mathbf{y}}_l. \quad (6.52)$$

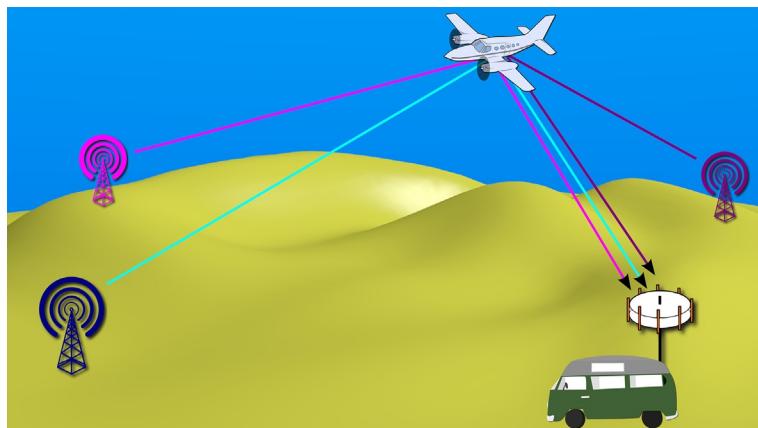
For the optimal case the number of echoes is  $L = M \cdot K$ , where  $M$  the number of transmitters and  $K$  the number of targets. However, in real scenarios  $L$  will be less than  $M \cdot K$  due to shadowing effects and Doppler zero constellations.

Let us assume a constellation as depicted in Fig. 6.24, which consists of three DVB-T stations, one receiver and a countable number of targets.

#### 6.5.4.1 Signal model

The presented model of this paper is a simplified model of the reality. It is assumed that targets are only maneuvering in x/y-direction, however, the third dimension can be easily incorporated. Let us assume that the scene is illuminated by  $M$  transmitters and observed by one receiver which are located at  $\mathbf{p}_{tm} = [x_{tm}, y_{tm}]^T$ ,  $m \in 1, \dots, M$  and  $\mathbf{p}_r = [x_r, y_r]^T$ , respectively. The  $k$ th target is described by its position vector  $\mathbf{p}_k = [x_k, y_k]^T$  and its speed vector  $\mathbf{v}_k = [v_{x_k}, v_{y_k}]^T$ .

The received signal is a composition of several time-delayed signals: strong direct signals from the transmitters ( $\tau_{m0}$ ), interlaced with echoes from static objects (which will be omitted as they are not of interest), and with echoes from  $K$  moving targets ( $\tau_{mk}$ ). Taking also some noise  $n(t)$  into account the received signal of the SFN receiver can be described by:



**FIG. 6.24**

---

Sensor network with DVB-T stations as transmitters and a single dedicated receiver.

$$\begin{aligned} y(t) &= \sum_{m=1}^M \alpha_{m0} x(t - \tau_{m0}) \\ &+ \dots + \sum_{m=1}^M \sum_{k=1}^K \alpha_{mk} x(t - \tau_{mk}) e^{-j2\pi(f_{D_{mk}} + f_c)\tau_{mk}} + n(t), \end{aligned} \quad (6.53)$$

where  $x(t)$  is the transmitted signal. The time delay for target  $k$  illuminated by transmitter  $m$  is:

$$\tau_{mk} = \frac{\|\mathbf{p}_k - \mathbf{p}_{tm}\| + \|\mathbf{p}_r - \mathbf{p}_k\|}{c} \quad (6.54)$$

and the Doppler shift:

$$f_{D_{mk}} = \frac{f_c}{c_0} (\mathbf{v}_k \cdot \mathbf{u}_{k\,tm} + \mathbf{v}_k \cdot \mathbf{u}_{k\,r}) \quad (6.55)$$

with  $\mathbf{p}_{tm, r, k}$  the position vectors,  $\mathbf{v}_k$  the velocity vector of target  $k$ , and  $\mathbf{u}_{k\,tm/k\,r}$  the direction vector pointing from target  $k$  to the transmitter  $m$  or to the receiver  $r$ , respectively.

For a distributed SFN network consisting of several DVB-T stations (Tx:  $\{1, \dots, m\}$ ) and one receiver the following relations exist:

$$\begin{aligned} \mathbf{y} &= \sum_{l=1}^L \tilde{\mathbf{y}}_l, \\ \mathbf{A} &= [\mathbf{A}_1, \dots, \mathbf{A}_m], \\ \mathbf{s} &= [\mathbf{s}_1^T, \dots, \mathbf{s}_m^T]^T, \end{aligned} \quad (6.56)$$

where  $\tilde{\mathbf{y}}_l$  the response of an extended target,  $\mathbf{A}_m$  the sensing matrix for the measurement combination of transmitter  $m$ , target, and receiver, and  $\mathbf{s}_m = [s_{m1}, \dots, s_{mP}]^T$  the target state vector. Groups are formed by  $[\mathbf{s}_1(1), \mathbf{s}_2(1), \mathbf{s}_3(1)]$ ,  $[\mathbf{s}_1(2), \mathbf{s}_2(2), \mathbf{s}_3(2)]$ , ...,  $[\mathbf{s}_1(K), \mathbf{s}_2(K), \mathbf{s}_3(K)]$ .

The estimation of the target state vector  $\mathbf{s}$  is then performed by an Extended Group Orthogonal Matching Pursuit (EGOMP) framework which takes into account the interlaced groups and the clustering of the extended targets, as described by Eq. (6.52):

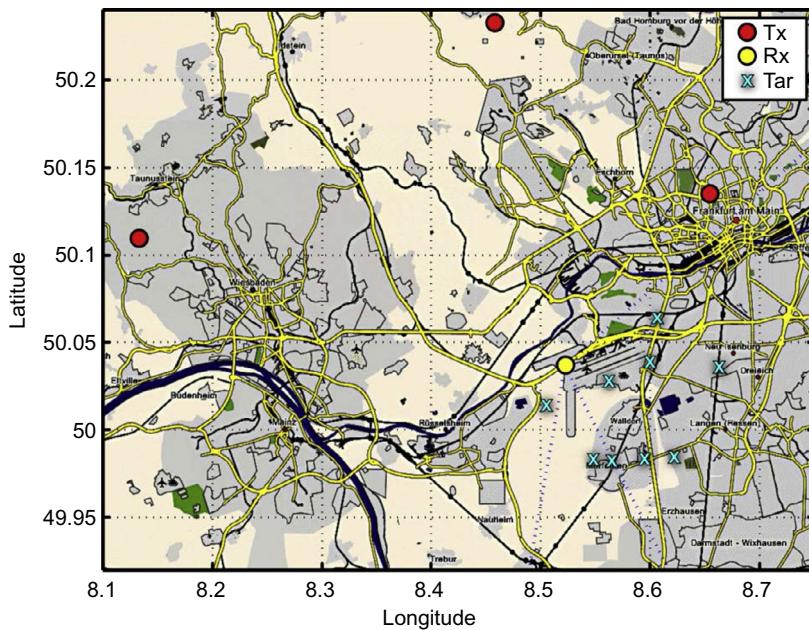
$$\min \sum_i \|s_{ai}\|_2 \text{ subject to } \|\mathbf{A}\mathbf{s} - \mathbf{y}\|_2 \leq \epsilon, \quad (6.57)$$

where  $\alpha_i$  relates to the group  $i$  and  $\epsilon$  is a threshold determined by the present noise.

#### 6.5.4.2 Verification

To verify the presented approach the following setup is considered: three DVB-T stations illuminate the scene which is observed by one receiver. The scene consists of several maneuvering targets, noted in Fig. 6.25 by cyan crosses (gray crosses in the print version).

To increase the detection probability we integrate over a large number of OFDM symbols. To preprocess the data we assume that the broadcast signal is a narrow band

**FIG. 6.25**

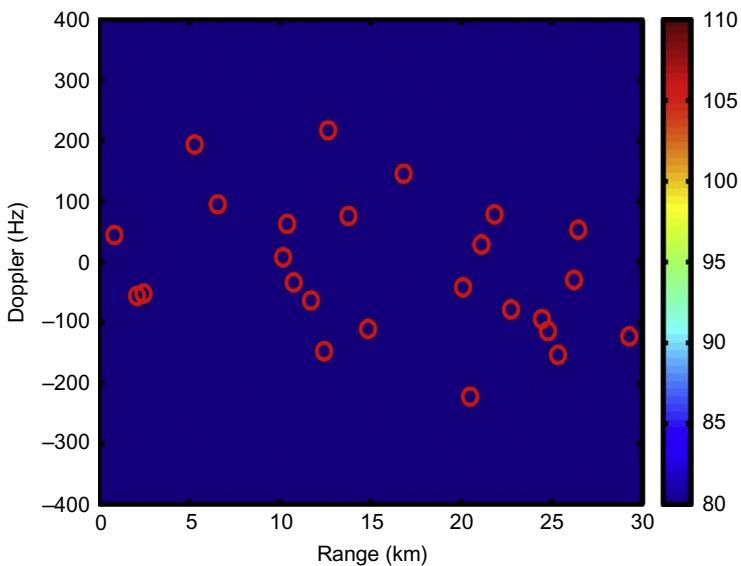
Single frequency surveillance radar network with three DVB-T stations (red dots) and one receiver (yellow dot) located near airport Frankfurt.

signal. With this boundary condition and if the target movement is less than the range resolution ( $\delta r_{target} \ll \Delta R_m$ ) for each OFDM-symbol length ( $T_{symbol}$ ) the Doppler frequency modulation can be approximated by a constant phase shift  $f_D = \zeta f_c$ . Due to these facts fast moving objects are smeared out in the range/Doppler map.

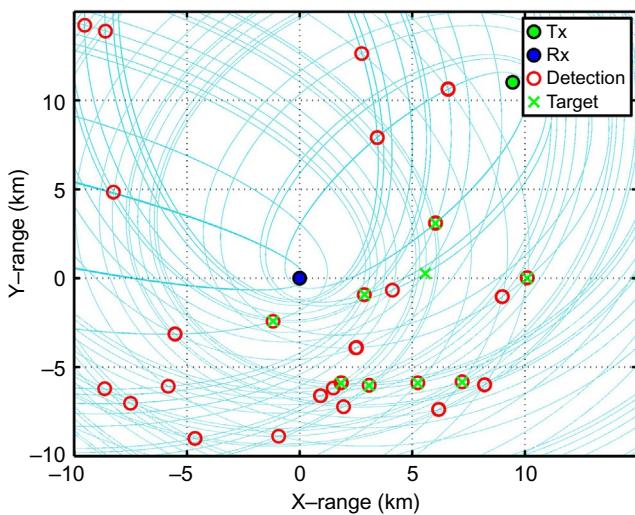
[Fig. 6.26](#) shows the corresponding range/Doppler plot where circles represent the detections after the CFAR detection stage.

Applying the traditional data fusion technique, where each detection is transformed to an ellipse around the corresponding tx/rx-pair, the detections convert into the result shown in [Fig. 6.27](#). A target position corresponds to the intersection of three ellipses. For a single target the problem is solvable, because only three ellipses are generated. However, if the target number is higher these ellipses generate more intersections than existing objects. This is shown in [Fig. 6.27](#) where red circles (dark gray circles in print versions) mark possible target positions. The green crosses (gray crosses in print versions) identify the correct positions of the simulated targets.

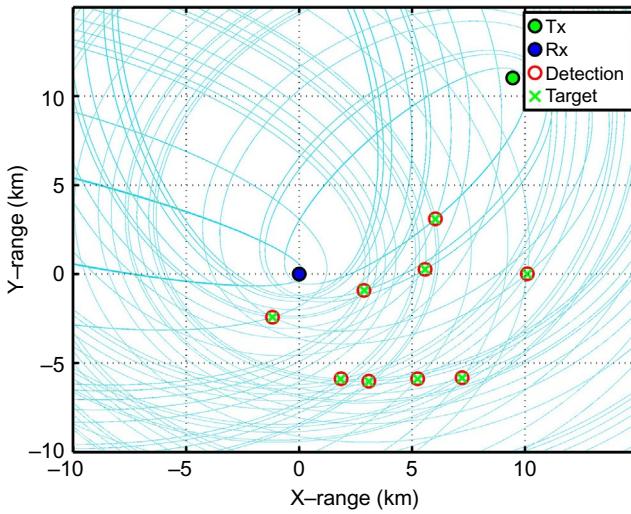
In comparison to the traditional approach the data fusion method described by Eq. (6.57) detects only the existing targets, as shown in [Fig. 6.28](#).

**FIG. 6.26**

Range/Doppler plot of the simulation. Detections are marked by circles.

**FIG. 6.27**

Detection result of the traditional data fusion approach in the Cartesian coordinate system. The red circles (dark gray circles in print versions) show the estimated targets positions and the green crosses (gray crosses in print versions) the correct positions.

**FIG. 6.28**

Detection result of the EGOMP method in the Cartesian coordinate system. The green crosses (gray crosses in print versions) denotes the positions of the simulated targets.

## 6.6 CONCLUSION

The intention of this paper has been to describe how the new theory of sparse sensing technique can be applied to radar systems. This technique is applicable to a sparse observation environment. It offers the ability to extract information of a target without fulfilling the Nyquist-Shannon criteria. The performance of the overall system is not decreased. Investigations have shown that detection and estimation performance depend strongly on the knowledge of the involved transmit/receive nodes and of the environment. If these points can be exactly modeled by the sensing matrix sparse sensing techniques are able to recover the concealed target information from less measurements.

One should keep in mind that sparse sensing theory assumes that target parameters can be described on a discrete parameter grid and depend on several system parameters. For instance, the resolution of this grid in range is determined by the frequency bandwidth and the Doppler by the coherent integration time. Targets not located on the search grid cause side-lobes in the neighboring grid points. To overcome this problem one possible strategy is to expand the search field of CS by a second derivation of the reference signal, as proposed by [46]. This second derivation provides information for adjusting the search grid to grid points which correspond to target parameters.

Investigations have shown that CS techniques are able to increase the resolution by a factor of two. Going beyond this point the column correlation of the sensing

matrix, as described by Herman and Stohmer [10], rises and the CS algorithm fails to distinguish between two different target states.

It has to be mentioned that compressive sensing/sparse reconstruction technique is no wizardry to extract concealed information from a noisy data or to recover signals covered by sidelobes caused by other targets.

---

## REFERENCES

- [1] J. Schwartz, B. Steinberg, Ultrasparse, ultrawideband arrays, *IEEE Trans. Ultrason. Ferroelectr. Freq. Control* 45 (2) (1998) 376–393.
- [2] J.-F. Hopperstad, S. Holm, The coarray of sparse arrays with minimum sidelobe level. in: Proc. IEEE NORSIG-98, 1998, pp. 137–140, <https://doi.org/10.1109/10.1.148.4618>.
- [3] E. Fishler, A. Haimovich, R. Blum, L. Cimini, D. Chizhik, R. Valenzuela, MIMO radar: an idea whose time has come, in: *IEEE Radar Conference 2004*, 2004, pp. 71–78.
- [4] D. Donoho, Compressed Sensing, 2004, <http://sys.cs.pdx.edu/trac/syn/export/36/CCS/related%20work/CompressedSensing091604.pdf> (Accessed 14 September 2004).
- [5] D. Wipf, B. Rao,  $l_0$ -norm minimization for basis selection, *Adv. Neural Inf. Process. Syst.* 17 (2005) 1513–1520.
- [6] D. Donoho, Compressed sensing, *IEEE Trans. Inf. Theory* 52 (4) (2006) 1289–1306.
- [7] E. Candès, J. Romberg, T. Tao, Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information, *IEEE Trans. Inf. Theory* 52 (2006) 489–509.
- [8] R. Baraniuk, Compressive sensing, *IEEE Signal Process. Mag.* 24 (4) (2007) 118–121.
- [9] E. Candès, M. Wakin, An introduction to compressive sampling, *IEEE Signal Process. Mag.* (2008) 21–30.
- [10] M. Herman, T. Stohmer, High-resolution radar via compressed sensing, *IEEE Trans. Signal Process.* 57 (6) (2008) 2275–2284, <https://doi.org/10.1109/TSP.2009.2014277>.
- [11] P. Gill, A. Wang, A. Molnar, The in-crowd algorithm for fast basis pursuit denoising, *IEEE Trans. Signal Process.* 59 (10) (2011) 4594–4605, <https://doi.org/10.1109/TSP.2011.2161292>.
- [12] S. Mallat, Z. Zhang, Matching pursuits with time-frequency dictionaries, *IEEE Trans. Signal Process.* 41 (12) (1993) 3397–3415, <https://doi.org/10.1109/78.258082>.
- [13] D. Needell, J. Tropp, CoSaMP: iterative signal recovery from incomplete and inaccurate samples, in: *Information Theory and Applications*, 31 January 2008, San Diego, 2008, pp. 1–25, arXiv:0803.2392.
- [14] E. van den Berg, M.P. Friedlander, Probing the Pareto frontier for basis pursuit solutions *SIAM J. Sci. Comput.* 32 (2) (2008) 890–912, <https://doi.org/10.1137/080714488>.
- [15] R. Baraniuk, P. Steeghs, Compressive radar imaging, in: *IEEE Radar Conference*, Waltham, MA, 2007, pp. 128–133.
- [16] E. Candès, T. Tao, Near optimal signal recovery from random projections: universal encoding strategies? *IEEE Trans. Inf. Theory* 52 (2006) 5406–5425.
- [17] E. Candès, J. Romberg, T. Tao, Stable signal recovery from incomplete and inaccurate measurements, *Commun. Pure Appl. Math.* 59 (8) (2006) 1207–1223.
- [18] J. Haupt, R.D. Nowak, Signal reconstruction from noisy random projections, *IEEE Trans. Inf. Theory* 52 (9) (2006) 4036–4048.

- [19] D. Donoho, M. Elad, V. Temlyakov, Stable recovery of sparse overcomplete representations in the presence of noise, *IEEE Trans. Inf. Theory* 52 (1) (2006) 6–18.
- [20] R. Baraniuk, M. Davenport, R. DeVore, M. Wakin, A simple proof of the restricted isometry property for random matrices. *Constr. Approx.* 28 (3) (2008) 253–263, <https://doi.org/10.1007/s00365-007-9003-x>.
- [21] H. Rauhut, Compressive sensing and structured random matrices, in: *Theoretical Foundations and Numerical Methods for Sparse Recovery*, Radon Series on Computational and Applied Mathematics, deGruyter, Berlin, 2010, pp. 1–92.
- [22] S. Foucart, H. Rauhut, *A Mathematical Introduction to Compressive Sensing*. Springer Science+Business Media, New York, 2013, ISBN 978-0-8176-4947-0, <https://doi.org/10.1007/978-0-8176-4948-7>.
- [23] D. Donoho, M. Elad, Optimally sparse representation in general (nonorthogonal) dictionaries via  $\ell_1$  minimization, *Proc. Natl. Acad. Sci. U. S. A.* 100 (5) (2003) 2197–2202.
- [24] A. Cohen, W. Dahmen, R. DeVore, Compressed sensing and best  $k$ -term approximation, *J. Am. Math. Soc.* 22 (2009) 211–231.
- [25] J.A. Tropp, A.C. Gilbert, Signal recovery from partial information via orthogonal matching pursuit, *IEEE Trans. Inf. Theory* 53 (12) (2007) 4655–4666.
- [26] R.A. Rankin, The closest packing of spherical caps in  $n$  dimensions, *Proc. Glasgow Math. Assoc.* 2 (1955) 139–144.
- [27] L.R. Welch, Lower bounds on the maximum cross-correlation of signals. *IEEE Trans. Inf. Theory* 20 (3) (1974) 397–399, <https://doi.org/10.1109/TIT.1974.1055219>.
- [28] M. Mishali, Y.C. Eldar, Xampling: compressed sensing of analog signals, in: Y.C. Eldar, G. Kutyniok (Eds.), *Compressed Sensing: Theory and Applications* (Chapter 3), ISBN 978-1-107-00558-7.
- [29] O. Bar-Ilan, Y.C. Eldar, Sub-Nyquist radar via Doppler focusing, CoRR, vol. abs/1211.0722, 2012 [Online]. Available at: <http://arxiv.org/abs/1211.0722>
- [30] D. Cohen, Y.C. Eldar, Reduced time-on-target in pulse Doppler radar: slow time domain compressed sensing, in: *IEEE Radar Conference*, Washington, DC, 2016.
- [31] T. Wimalajeewa, Y.C. Eldar, P.K. Varshney, Recovery of sparse matrices via matrix sketching, CoRR, vol. abs/1311.2448, 2013 [Online]. Available at: <http://arxiv.org/abs/1311.2448>.
- [32] L.H. Nguyen, T.T. Do, T.D. Tran, Sparse model and sparse recovery with ultra-wideband SAR applications, in: 1st International Workshop on Compressed Sensing Applied to Radar (CoSeRa 2012), 2012, <http://workshops.fhr.fraunhofer.de/cosera>.
- [33] J.H.G. Ender, On compressive sensing applied to radar. *Signal Process.* 90 (5) (2010) 1402–1414, <https://doi.org/10.1109/ICASSP.2008.4518185>.
- [34] D.H. Johnson, D.E. Dudgeon, *Array Signal Processing: Concepts and Techniques*, Prentice Hall, Englewood Cliffs, NJ, 1993.
- [35] A. Gretistas, M.D. Plumley, A multichannel spatial compressed sensing approach for direction of arrival estimation, in: *Latent Variable Analysis and Signal Separation*, Lecture Notes in Computer Science, vol. 6365, 2010, pp. 458–465.
- [36] P. Berardino, G. Fornaro, R. Lanari, E. Sansosti, A new algorithm for surface deformation monitoring based on small baseline differential SAR interferograms, *IEEE Trans. Geosc. Remote Sens.* 40 (11) (2002) 2375–2383.
- [37] A. Reigber, A. Moreira, First demonstration of airborne SAR tomography using multi-baseline L-band data, *IEEE Trans. Geosc. Remote Sens.* 38 (5) (2000) 2142–2152.

- [38] G. Fornaro, F. Serafino, F. Soldovieri, Three-dimensional focusing with multipass SAR data, *IEEE Trans. Geosc. Remote Sens.* 41 (3) (2003) 507–514.
- [39] F. Lombardini, Differential tomography: a new framework for SAR interferometry, *IEEE Trans. Geosc. Remote Sens.* 43 (1) (2005) 37–44.
- [40] G. Fornaro, F. Lombardini, F. Serafino, Multidimensional imaging with ERS-data, in: Proc. Fringe Workshop, Frascati, Italy, 28 Nov.–02 Dec., 2005, 2005.
- [41] X. Zhu, R. Bamler, Very high resolution spaceborne SAR tomography in urban environment, *IEEE Trans. Geosc. Remote Sens.* 48 (12) (2010) 4296–4308.
- [42] X. Zhu, R. Bamler, Super-resolution power and robustness of compressive sensing for spectral estimation with application to spaceborne tomographic SAR, *IEEE Trans. Geosc. Remote Sens.* 450 (1) (2012) 247–258.
- [43] C. Baker, An introduction to multistatic radar, in: NATO SET-136 Lecture Series “Multistatic Surveillance and Reconnaissance: Sensor, Signals and Data Fusion”, 2009.
- [44] M. Weiß, Multi-sensor systems: multiplicity helps, in: RTO-EN-SET-157-2010—Multi-sensor Fusion: Advanced Methodology and Applications, ISBN 978-92-837-0114-9.
- [45] D. O'Hagan, M. Ummenhofer, H. Kuschel, J. Heckenbach, A passive/active dual mode radar concept, in: 14th International Radar Symposium IRS 2013I, Dresden, pp. 136–142.
- [46] J.H.G. Ender, A compressive sensing approach to the fusion of PCL sensors, in: Proceedings of the 21st European Signal Processing Conference (EUSIPCO), 9–13 Sept. 2013, ISBN 978-1-4799-3687-8, ISBN 978-0-9928626-0-2, pp. 555–559.
- [47] X. Lv, G. Bi, C. Wan, The group lasso for stable recovery of block-sparse signal representations, *IEEE Trans. Signal Process.* 59 (4) 1371–1382.
- [48] Y. Eldar, P. Kuppinger, H. Bolcskei, Block-sparse signals uncertainty relations and efficient recovery, *IEEE Trans. Signal Process.* 58 (6) (2010) 3042–3054.
- [49] M. Weiß, Passive WLAN radar network using compressive sensing technique. *IET Radar Sonar Navig.* 9 (1) (2015) 84–91, <https://doi.org/10.1049/iet-rsn.2014.0073>.
- [50] F. Colone, P. Falcone, C. Bongianni, P. Lombardo, Wifi-based passive bistatic radar, data processing schemes and experimental results, *IEEE Trans. Aerosp. Electron. Syst.* 48 (2) (2012) 1061–1079.
- [51] ETSI Standard: EN 300 744 V1.5.1, Digital Video Broadcasting (DVB); framing structure, channel coding and modulation for digital terrestrial television, Available at: <https://web.archive.org/web/20131023081351/http://pda.etsi.org/pda/queryform.asp>.

---

## FURTHER READING

- Y. Eldar, G. Kutyniok, *Compressed Sensing: Theory and Applications*, Cambridge University Press, Cambridge, 2012.  
<http://www.seti.org/ata>, <http://archive.seti.org/pdfs/Shostak-spring2009-EnS.pdf>.
- A.C. Gürbüz, J.H. McClellan, V. Cevher, A compressive beamforming method. in: ICASSP 2008, IEEE International Conference on Acoustics, Speech and Signal Processing, 2008, pp. 2617–2620, <https://doi.org/10.1109/ICASSP.2008.4518185>.
- A. Ferretti, C. Prati, F. Rocca, Nonlinear subsidence rate estimation using the permanent scatterers in differential SAR interferometry, *IEEE Trans. Geosc. Remote Sens.* 38 (5) (2000) 2202–2212.
- A. De Maio, G. Fornaro, A. Pauciullo, Detection of single scatterers in multidimensional SAR imaging, *IEEE Trans. Geosci. Remote Sens.* 47 (7) (2009) 2284–2297.

- A. Pauciullo, D. Reale, A. De Maio, G. Fornaro, Detection of double scatterers in SAR tomography, *IEEE Trans. Geosci. Remote Sens.* 50 (9) (2012) 3567–3586.
- G. Fornaro, F. Lombardini, F. Serafino, Three-dimensional multipass SAR focusing: experiments with long-term spaceborne data, *IEEE Trans. Geosci. Remote Sens.* 43 (4) (2005) 702–714.
- M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, *J. R. Stat. Soc. Ser. B* 68 (1) (2007) 49–67.
- T. Fawcett, ROC graphs: notes and practical considerations for researchers, *Mach. Learn.* 31 (2004) 1–38.
- G.W. Stimson, H. Griffiths, C. Baker, D. Adamy, *Stimson's Introduction to Airborne Radar*, third ed., 2014, ISBN 978-1-61353-022-1.
- B. Cook, *Radar Signals: An Introduction to Theory and Application*, Artech House, Norwood, MA, 1993.
- X. Zhu, R. Bamler, Demonstration of super-resolution for tomographic SAR imaging in urban environment, *IEEE Trans. Geosci. Remote Sens.* 50 (8) (2012) 3150–3157.

# Millimeter-wave integrated radar systems and techniques

# 7

Akram Al-Hourani<sup>\*</sup>, Robin J. Evans<sup>†</sup>, Peter M. Farrell<sup>†</sup>, Bill Moran<sup>\*</sup>,  
Marco Martorella<sup>‡</sup>, Sithamparanathan Kandeepan<sup>†</sup>,  
Stan Skafidas<sup>†</sup>, Udaya Parampalli<sup>†</sup>

*RMIT University, Melbourne, VIC, Australia* <sup>\*</sup>*University of Melbourne, Melbourne, VIC, Australia*  
<sup>†</sup>*University of Pisa, Pisa, Italy*  
<sup>‡</sup>

## 7.1 INTEGRATED RADAR: TRENDS AND CHALLENGES

### 7.1.1 SYSTEM DESIGN CHALLENGES: SIZE AND COST

While the essential idea of radar is very simple, its significance in applications is such that the ongoing effort to improve the performance and capabilities has resulted in radar being a major technology driver since its inception. It is anticipated that this trend is set to continue in an accelerated manner given the emerging opportunities in millimeter-wave radar-on-a-chip made possible by Complementary Metal Oxide Semiconductors (CMOS) technology scaling (Moore's law) and recent advances in adaptive waveform design and scheduling.

Interest in consumer applications of radar is of course not new. In 1993 the IEE held a symposium on this very topic [2] and there are already small low-cost hand held Doppler radar systems on the market (e.g., Pocket Radar and Google's project Soli [3]), there are also other examples in automotive industry.

Automotive radar is one of the current challenges driving innovation in small low-cost integrated millimeter-wave radar. Another promising application concerns navigation and sensing systems for unmanned aerial vehicles (UAVs). The ready availability of tiny, cheap, radar systems is likely to open up vast new applications similar to the transformations arising from the availability of tiny and cheap GPS systems. The ongoing electronic technology scaling is providing the opportunity to move to higher radio frequencies (RFs) hence enabling complete radar systems (RF and the digital signal processor) to be integrated onto a single chip, where the essential electromagnetic (EM) interference shielding between the RF and the DSP parts becomes a critical issue. Smart new waveform diversity and scheduling techniques [4] together with innovations in small antenna technology, mean advanced consumer radar systems will become a reality in the near future.

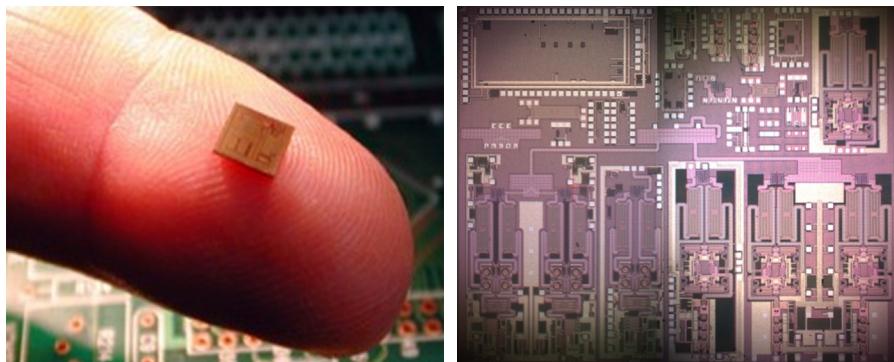
### 7.1.2 SINGLE CHIP RF SYSTEM

Many research groups around the world are developing different types of *single chip* radar systems mostly focusing on automotive radar applications. For an overview of these activities see, for example, Refs. [5–8]. In this brief introduction we provide an overview of the radar-on-a-chip system developed at the University of Melbourne. The  $5\text{ mm} \times 5\text{ mm}$  CMOS chip shown in Fig. 7.1 is built using 65 nm technology.

The chip itself contains a full RF transceiver system operating at 76–77 GHz. As seen in Fig. 7.2 a typical configuration consists of two transmit chains and four receive chains. All required components including passives, amplifiers, mixers, and oscillators are contained on the same chip. A transmit power of around 10 dBm at 77 GHz has been achieved [9]. A receiver sensitivity of approximately  $-100\text{ dBm}$  for an output signal-to-noise ratio (SNR) of 10 dB has been measured [7]. It is important to note that the sensitivity requirement varies depending on the application such as the maximum speed and the maximum range. The ability to utilize multiple transmitter and receiver chains means that a radar can simultaneously sense different spatial scenarios, such as different feeds from different antenna elements, or different polarizations. This is similar to multi-input-multi-output communication technique that can increase the diversity and capacity of a communication system. Multiple Rx chains also support antenna array processing.

### 7.1.3 ANTENNA SYSTEMS

In order to meet the size and cost requirements of a consumer radar, patch antennas can be used with low-cost lens technology [10]. The lens can be in the form of a dielectric dome or metallic plates [10]. The antenna shown in Fig. 7.3 consists of a serial patch structure with a low-cost dielectric lens. The performance at 77 GHz is shown in Fig. 7.4 depicting the elevation and azimuth radiation patterns with and without lens technology.



**FIG. 7.1**

A photograph depicting a  $5\text{ mm} \times 5\text{ mm}$  radar RF chip built with 65 nm technology [1], designed in the University of Melbourne.

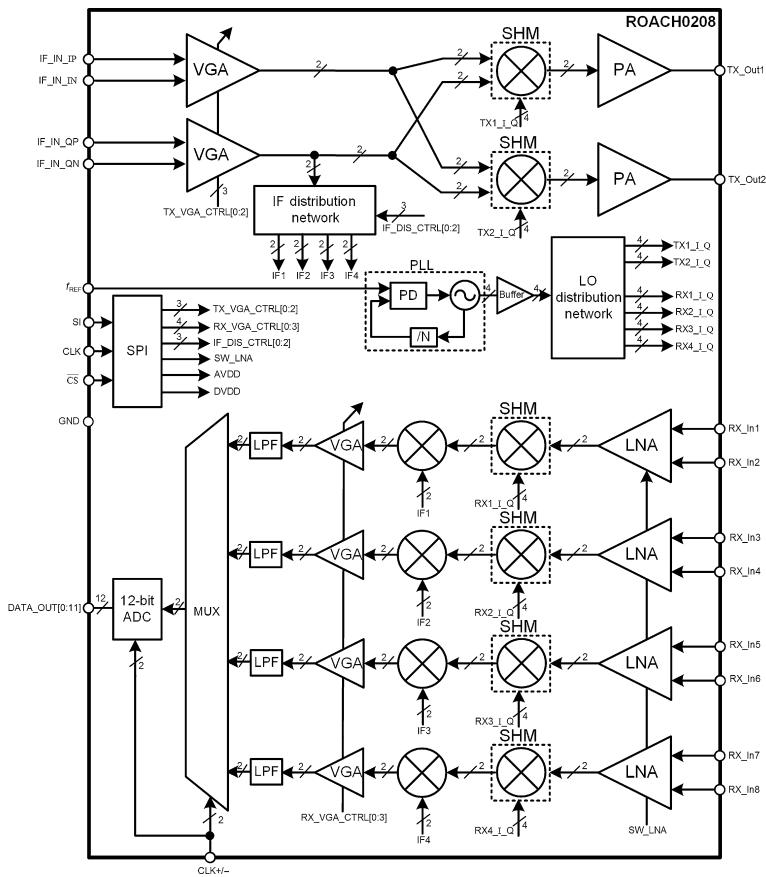


FIG. 7.2

A block diagram for radar consists of two transmit chains and four receive chains [1].

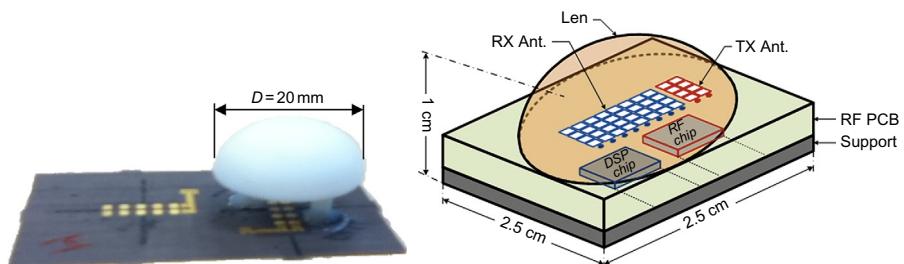
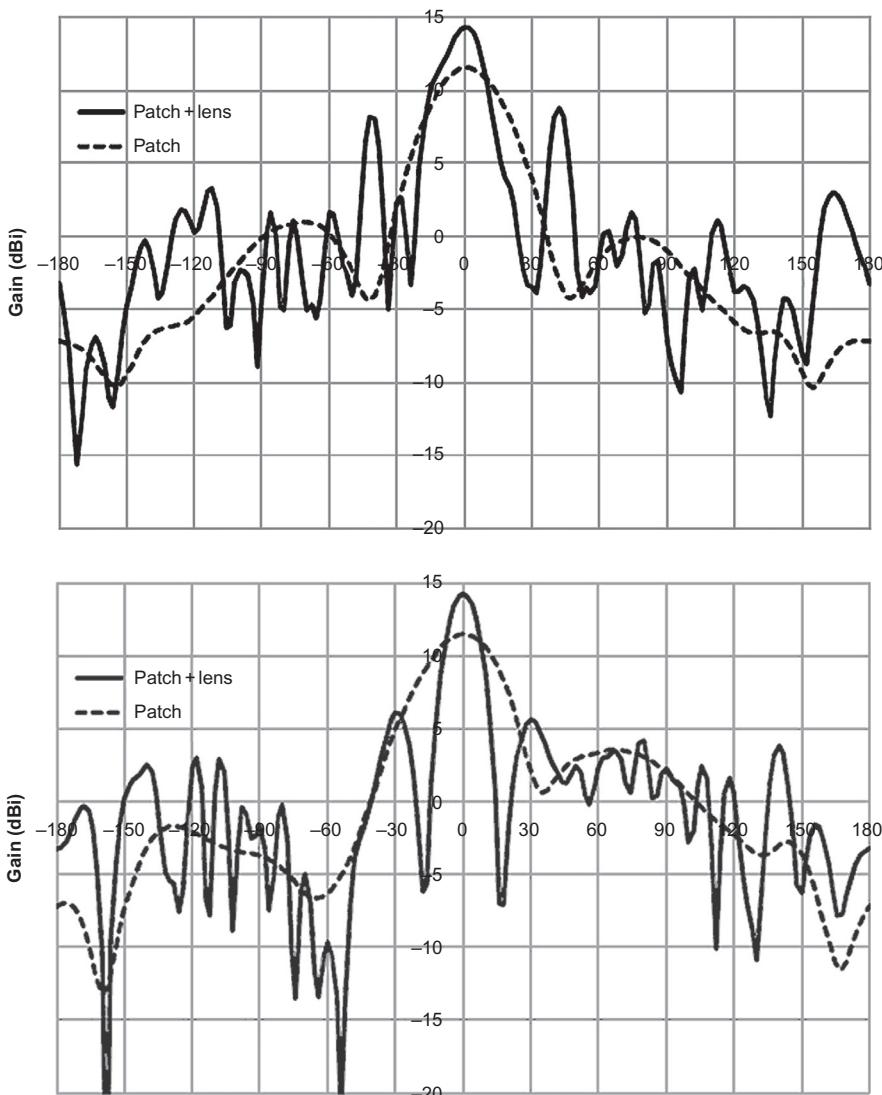


FIG. 7.3

(Left) A photograph of the patch antenna with one lens mounted on the transmitter array.

(Right) An illustration showing the packaging of the radar on chip [11].

**FIG. 7.4**

A comparison between the antenna gain with and without the effect of the lens [10].

#### 7.1.4 INTERFERENCE CHALLENGES

Given this upcoming vast deployment of consumer radars, it is anticipated that significant radar-to-radar interference will arise due to the shared spectrum use and unavoidable proximity. For example, in typical automotive applications, a radar can be easily blinded or confused by vehicles traveling in the opposite direction

resulting in degraded performance in radar detection ability that might coincide with a split-second critical road situation. Similarly, backward looking radars can interfere with forward looking radars for vehicles traveling in the same direction. The interference is largely caused by the use of shared spectrum and the inherent lack of coordination between radars resulting from the lack of centralized control and resource allocation. There exist, of course, many tools to handle radar interference including clever waveform design, fast adaptive antenna methods particularly nulling, polarization switching, various signal processing methodologies, and many more [12].

Thus there is now a growing activity aimed at understanding and addressing the problem of mutual interference arising from overlapping radar signals. Many of these attempts have been initiated by the automotive industry such as the EU project MOSARIM [13], which investigated automotive radar interference by conducting experimental road measurements and by conducting complex ray-tracing simulations. This particular project also explored some possible interference mitigation techniques. An important conclusion from this project suggests that (particularly for linear frequency modulated [LFM] radar waveforms) interfering radars are unlikely to cause ghost targets but rather they will create noise-like combined interference. Ghost targets due to interference were studied analytically in [14] and were observed when two identical radars are utilized with identical waveforms (but not synchronized). However, the same paper found that it is more likely that the radars will cause a noise-like interference for typical practical scenarios. Exploiting this observation, refs. [15, 16] suggest a practical approach for randomizing chirp sweep frequency in order to guarantee noise-like interference, thus aiming to reduce the false alarm probability caused by ghost targets. Random stepped frequency (RSF) radar is also suggested in the literature to mitigate radar-to-radar interference such as the work in [17] suggesting that RSF would also suppress range ambiguity and enhance covert detection. Practical system algorithms to efficiently implement RSF in automotive applications can be found in the patents [18, 19] suggesting reduced interference when utilizing this scheme. We shall see in [Section 7.4](#) a tractable approach to characterize the arising interference using tools from stochastic geometry.

### 7.1.5 AUTOMOTIVE RADAR: TRENDS AND STANDARDIZATION EFFORTS

As an important application of CMOS radar, automotive radar is emerging as a key technology enabling intelligent and autonomous features in modern vehicles such as relieving drivers from monotonous tasks, reducing driver stress, and adding life-saving automatic interventions. Recently, automotive radar is implemented in many high-end cars to enable essential safety and comfort features including adaptive cruise control and automatic emergency breaking systems where a vehicle can steeply decelerate without driver involvement to avoid a potential collision. The deployment of these features has thus far been limited to high-end vehicles because

of the high cost of sensing technology. However, as discussed earlier, this situation is now rapidly changing and the expected vast global market penetration of automotive radar technology has required both international and local regulatory authorities to work in conjunction with the automotive industry to develop appropriate and harmonized standards. It is anticipated that by 2030 the penetration of automotive radars will reach around 65% in Europe and 50% in the United States as described in the International Telecommunication Union (ITU) document [20]. In another recommendation document [21], ITU classifies automotive radar into two main categories according to their ranging capabilities and safety requirements:

- *Category 1*: Designed for long distances up to 250 m, serving the adaptive cruise control and collision avoidance systems.
- *Category 2*: Designed for short and medium distances up to 50–100 m depending on the application, utilized for lane change assistance and rear traffic crossing alert.

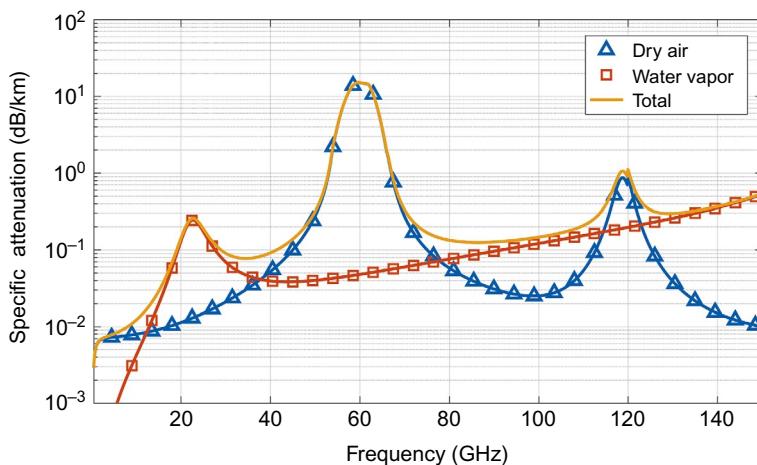
The bandwidth requirement of the long-range category is planned to be 1 GHz, with maximum allowed equivalent isotropic radiated power (EIRP) of 43 dBm [22, 23], where the achievable EIRP by integrated radars is much lower. The medium/short-range category has less power allowance, and a wider spectrum bandwidth to support higher range resolution for close targets, based on the typical relation of resolution and bandwidth [15],  $\Delta R = \frac{c}{2B}$ , where  $\Delta R$  is the range resolution,  $c$  is the speed of light, and  $B$  is the used bandwidth.

## 7.2 CHANNEL MODELING FOR MILLIMETER-WAVE RADAR

The EM energy transmitted by the radar travels in a propagation environment that adds extra losses on top of the natural wavefront expansion. These losses are caused by atmospheric attenuation and absorption, and by the various reflectors, such as small airborne particles, snow and rain, which all cause unwanted scattering and clutter. In this section we discuss the specific nature of the millimeter-wave radar channel.

### 7.2.1 PROPAGATION PROPERTIES IN MILLIMETER-WAVE

The spectrum range between 30 and 300 GHz is referred to as the millimeter-wave spectrum since the wavelengths for these frequencies are in an order of millimeters (less than 10 mm). Many favorable traits in radar systems can leverage the millimeter-wave spectrum such as small components size, large availability of bandwidth, and low mutual interference between radars [12]. High directivity can be achieved in small antennas, thus reducing the demand on high-power amplifiers, making millimeter-wave a suitable choice for CMOS applications.

**FIG. 7.5**

Atmospheric absorption due to water vapor and gases in the millimeter-wave spectrum, the graph is generated based on ITU guidelines [24].

RF propagation is usually impaired by atmospheric absorption caused by the resonance of water vapor molecules and oxygen gas molecules, as shown in Fig. 7.5, which is based on ITU guidelines, with a total atmospheric absorption of 0.23, 13.55, and 0.15 dB/km at frequencies 24, 61, and 76.5 GHz, respectively, at 25°C, 50% relative humidity, and a pressure of 101.3 kPa [24]. In general, the signals using millimeter-wave spectrum endure higher atmospheric absorption than signals using lower frequencies. Also, heavy rain can cause relatively higher attenuation in the millimeter-wave spectrum due to raindrop size being comparable with the wavelength. However, these impairments do not play a major factor for consumer short-range radars, where target distances are an order of 300 m. Below this range, absorptions sum to less than 0.1 dB for a  $2 \times 300$  m total return path under clear weather conditions. However, rain will contribute an additional 3–15 dB depending on the rain rate varying from light rain (5 mm/h) to heavy rain (50 mm/h) [25].

### 7.2.2 MILLIMETER-WAVE RADAR EQUATION

The signal transmitted by a radar and reflected from a target (or targets) is well-characterized in the literature to follow the *radar equation* [12]. To improve the accuracy of the channel model, it is common to include additional factor to account for losses such as atmospheric absorption as discussed in Section 7.2.1 as a distance dependent factor  $\eta$ . Thus the modified radar equation takes the form

$$S = \underbrace{\frac{P_o \eta G_t}{4\pi R^2}}_{\text{Incident signal}} \times \underbrace{\frac{\sigma_c}{4\pi R^2} A_e}_{\text{Reflected signal}} = \gamma_1 \gamma_2 P_o R^{-4} \eta, \quad (7.1)$$

which models the returned signal power, where  $P_o$  is the radar transmit power;  $R$  is the target range, that is, the distance to the target;  $G_t, A_e$  are the antenna gain and its effective area, respectively; and  $\sigma_c$  is the radar cross-section (RCS) area of the target. The parameters  $\gamma_1$  and  $\gamma_2$  are given as

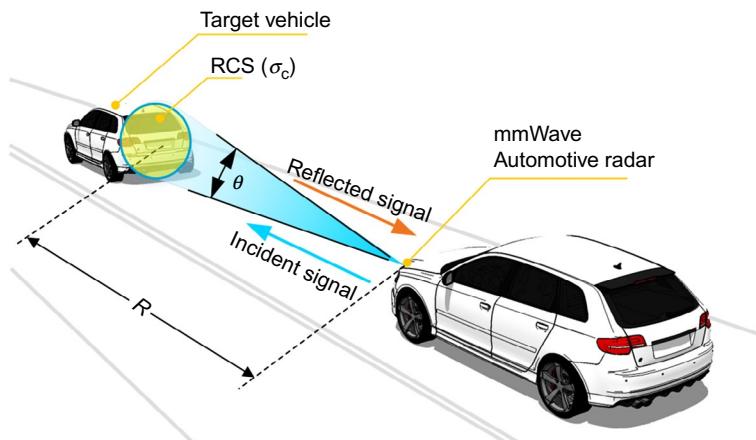
$$\gamma_1 = \frac{G_t A_e}{4\pi} = G_t^2 \left( \frac{c}{4\pi f} \right)^2, \quad (7.2)$$

$$\text{and } \gamma_2 = \frac{\sigma_c}{4\pi}, \quad (7.3)$$

where  $f$  is the operating frequency. We illustrate the parameters affecting a radar signal in Fig. 7.6, for millimeter-wave radar in the typical automotive scenario. See Ref. [12] for further discussions on the radar equation.

### 7.2.3 RAY TRACING FOR MILLIMETER-WAVE RADAR

Advances in computer-aided propagation tools based on ray-tracing simulations allow the prediction of signal strength in a very accurate manner. Ray-tracing simulation is based on the approximation of EM waves by optical rays. This approximation holds very well in the millimeter-wave region. The processes of EM ray tracing in radar involve two main steps; the first is to determine the geometrical paths that are emitted from the radar and reflected back, and the second is to determine the



**FIG. 7.6**

Illustration of some of the factors that affect the strength of the returned signal in typical automotive scenario.

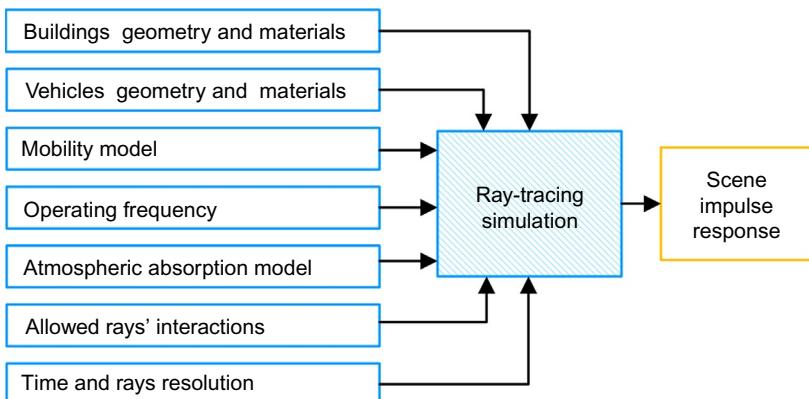
encountered interactions experienced by these rays (i.e., reflections, diffractions, and transmissions) [26]. In a time-varying environment such as an automotive radar situation, the rays will undergo a Doppler shift due to the movement of reflecting objects. This is usually handled in a postprocessing stage and not during the ray-tracing stage. Conducting ray tracing for automotive radars reduces the need for expensive field trials and provides a more controlled experimental environment. However, as with any other RF model, the accuracy lies in the stipulated mathematical and geometrical assumptions of the ray-tracing environment. These assumptions are listed following:

- the geometrical layout of the propagation environment; for example, layout of streets, buildings, vegetation, etc.;
- the resolution of the geometrical layout;
- the models of the EM properties of the objects;
- the number of allowable interactions (i.e., how many reflections, diffractions, and transmissions) is allowed for each ray. Taking into consideration that the computational complexity will exponentially increase for higher number of interactions.

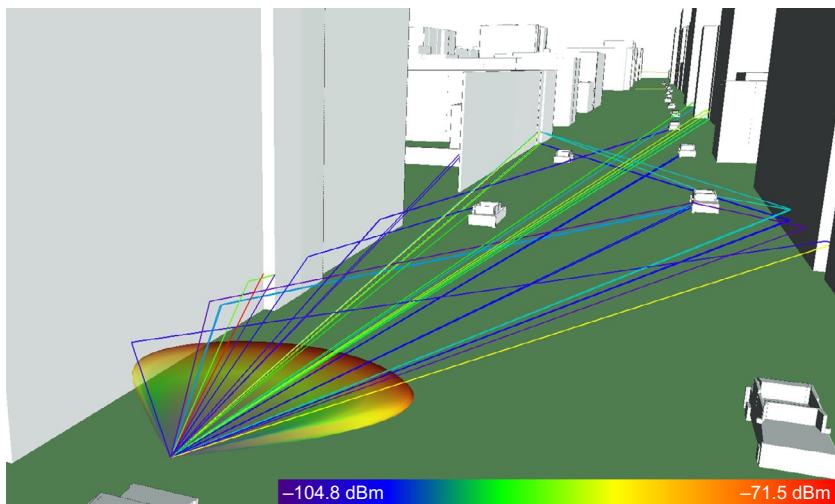
The process of ray-tracing simulation starts with the selection of the geometrical environment that is usually some subset from an urban 3D model. Simulating as many different environments as possible is preferable to verify the performance of automotive radars [27]. These environments are primarily grouped into the following [28]: highways, urban streets, suburban streets, and rural streets. Also, they are associated with a certain vehicle density and traffic velocity. After identifying the desired environments, a simplified database of objects is created for the fixed objects (buildings, vegetation, lamp posts, etc.) and for the mobile objects (vehicles, pedestrians, etc.). It is necessary to take into consideration that the database should include proper materials for each object. After preparing the database, ray-tracing can be run with a specified level of complexity based on the number of allowed interactions per ray, and the resolution of the time samples. A summary of the ray-tracing process is shown in Fig. 7.7.

A typical raw result of a ray-tracing simulator would be a list of the received rays stamped with the direction of arrival; direction of departure; and most importantly the delay, phase, and magnitude of the electric field. We depict in Fig. 7.8 a temporal snapshot of the rays reflected from buildings and other vehicles on the road. This simulation was performed at 77 GHz, with vertical polarization, an E-plane beamwidth of 4 degrees, and a H-plane beamwidth of 60 degrees. In this scenario, the vehicles on the left side of the road are driving forward at the same speed, while the vehicles on the right are driving in the opposite direction also at the same speed. The buildings are taken from an actual urban footprint database of the Carlton area in the city of Melbourne, Australia, near the University of Melbourne.

For time-varying scenarios, it is usually required to perform many simulations during a predefined time period. This provides an understanding of the time-varying nature of the scene including possible interferences arising from other radars. We depict in Fig. 7.9 a time series of the channel impulse responses (CIR) as seen by

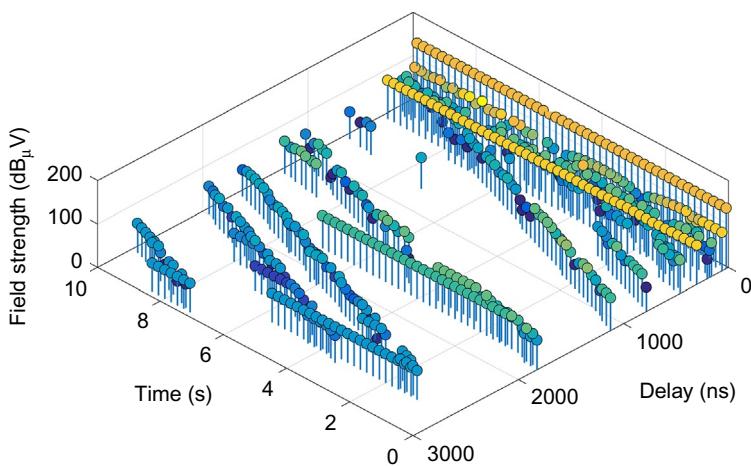
**FIG. 7.7**

A simplified process of millimeter-wave ray-tracing simulation.

**FIG. 7.8**

One temporal snapshot of radar ray tracing, showing the reflected-back rays and their strength.

the transmitting radar in [Fig. 7.8](#), where we note that the impulses with decreasing delays indicate approaching targets, while the impulses with fixed delays indicate the vehicles traveling at the same speed and in the same direction as the frame of reference associated with the vehicle under study. Doppler shifts do not directly appear in this time series CIR; however, they can be assigned by knowing the velocity of objects that rays have interacted with. This is usually performed by postprocessing the rays in a later stage.

**FIG. 7.9**

A 10-s sequence of impulse responses for the time-varying scene with the geometry given in Fig. 7.8.

#### 7.2.4 CLUTTER IN MILLIMETER-WAVE CMOS RADAR

*Clutter* is a core issue in radar systems. Clutter is caused by reflections from objects other than the targets of interest, usually referred to as *unwanted* reflections. It can cause major problems for detection and tracking of desired *targets* [29]. A typical example in consumer radar could be an automotive scenario, where the reflections from stationary (or slowly moving) pedestrians could be obscured, or overwhelmed, by background reflections from buildings or a rough road surface. Clutter can also arise from atmospheric effects, such as backscatter from airborne particles including dust, rain, snowfall, mist, etc. [30]. We often categorize clutter into two classes based on the resolution ability of the radar; *resolvable* and *unresolvable*. When radar resolution is sufficiently high, a small target can be detected even with the existence of nearby larger clutter reflections, provided that the sidelobes (in the ambiguity function) arising from clutter do not swamp the target. For example, it may be possible to separate targets and clutter in the Doppler domain even though they may not be separable in range and/or azimuth. In the second case, the radar is unable to resolve individual clutter scatterers, hence the return signal manifests itself in the receiver as noise [29], which can be characterized by its nonstationary statistical properties with both spatial and temporal variations. Modeling and empirical studies of unresolvable clutter have been carried out for various operating frequencies and radar waveforms involving different clutter situations; for example, sea clutter [31], ground clutter [32], weather clutter, etc. Some studies have been undertaken for millimeter-waves for certain waveforms [29]; however, much work needs to be done in this area.

The study of techniques to mitigate the negative effect on target detectability is extensively covered in a large range of approaches; for example, CFAR [33],

two-dimensional CFAR [34], adaptive techniques [12], etc. In addition to the range of existing techniques, a number of interesting new approaches have been tried such as augmenting a millimeter-wave radar with a vision system that can detect the geometrical characteristics of the scene (e.g., road border markings) and correlating these characteristics with the received clutter [35]. These vision systems might also be vulnerable to poor weather conditions such as mist and rain.

Phase noise is an important consideration in radar clutter processing as it deteriorates the ability to separate targets and clutter in the Doppler domain [33]. For short ranges, as it is often the case for CMOS radar, a significant correlation takes place between the phase noise that occurs during the transmission of the signal and the phase noise that occurs during the reception of the backscattered signal [36], thus mitigating the effect of LO phase noise in CMOS chip. This is the so-called *range correlation* effect.

## 7.3 WAVEFORM AND SIGNAL PROCESSING

Radar waveform design was put on a sound theoretical footing in 1953 when Philip Woodward introduced the radar ambiguity function [37], which was based on the matched-filter developed by Dwight North in 1943. The ambiguity function characterizes the performance of a matched-filter radar for any particular transmitted waveform. Unfortunately, however, it is not possible to precisely synthesize a transmit waveform based on a desired ambiguity function. Significant research effort has been devoted to the synthesis problem including the deep work of Wilcox [38] in 1960 on the group theoretic foundations of ambiguity theory that enabled synthesis of a certain restricted class of waveforms. A few years later, Hilbert-Schmidt operator approximation techniques were proposed by Sussman [39] and Vakman [40] to approximately synthesize waveforms with specified ambiguity function properties. See also Ref. [41] for recent work in this direction. Progress on this problem is difficult but it is deeply important and it has taken on a new flavor in recent years under names such as waveform diversity, adaptive waveforms, and waveform scheduling [4]. Digital waveform generation and fast adaptable digital matched-filter implementation mean that waveform diversity techniques can now be implemented even in low-cost single chip radar systems. The CMOS radar on a chip developed at Melbourne University utilizes adaptive digital matched filtering and scheduling of pseudo-random stepped frequency (PRSF) waveforms, to reduce clutter, mitigate against interference, and also to reduce computational load. Additional discussions on radar waveforms for CMOS applications are in the following sections.

### 7.3.1 TIME-BANDWIDTH PRODUCT AND RADAR RESOLUTION

Turin [42] found that regardless of the transmitted waveform the peak-signal to the average-noise ratio (SNR) of the matched-filter output is given by

$$\hat{\text{SNR}}_{\text{out}} = \frac{2E}{N_0}, \quad (7.4)$$

where  $E$  is the received signal energy and  $N_o$  is the noise power spectral density at the input of the matched filter. It can be shown [43], when the bandwidth of the waveform is small compared with the center frequency, that the *average* SNR at the output of the matched-filter is related to the SNR of the incoming signal as follows

$$\text{SNR}_{\text{out}} \approx T_1 B \times \text{SNR}_{\text{in}}, \quad (7.5)$$

where  $T_1$  is the waveform's total duration and  $B$  is its bandwidth. The bandwidth is usually taken at the  $-3$  dB point and, in the special case of a pulse waveform, the bandwidth is taken at the  $-4$  dB point. The quantity  $T_1 B$  is called the *time-bandwidth product*. The individual choice of  $B$  and  $T_1$  affects the radar resolution in the range and Doppler domains, respectively. Radar *range resolution* is defined as the minimum resolvable distance between two point-scatters. It can be shown that the approximate resolution of the delay  $\Delta\tau$  is equal to the inverse of the total bandwidth of the waveform [44],  $\Delta\tau \approx \frac{1}{B}$ . Accordingly the range resolution is

$$\Delta R \approx \frac{c}{2B}, \quad (7.6)$$

where  $c$  is the speed of light. Similarly, the *Doppler resolution* can be shown to be approximately equal to the inverse of the total duration of the waveform (i.e.,  $\Delta u \approx \frac{1}{T_1}$ ). Thus the velocity resolution is given by

$$\Delta v \approx \frac{c}{2T_1 f_o}, \quad (7.7)$$

where  $f_o$  is the carrier frequency.

A simple radar waveform is a sinusoidal pulse of short duration  $T_1$ , having an approximate bandwidth of  $B = \frac{1}{T_1}$ . Thus the time-bandwidth product of this simple pulse waveform is  $T_1(1/T_1) = 1$ , providing no SNR enhancement at the output of the matched-filter, as per Eq. (7.5). In order to enhance the SNR, and hence the ability of a radar to detect targets, we need waveforms that have much higher bandwidth than  $1/T_1$ .

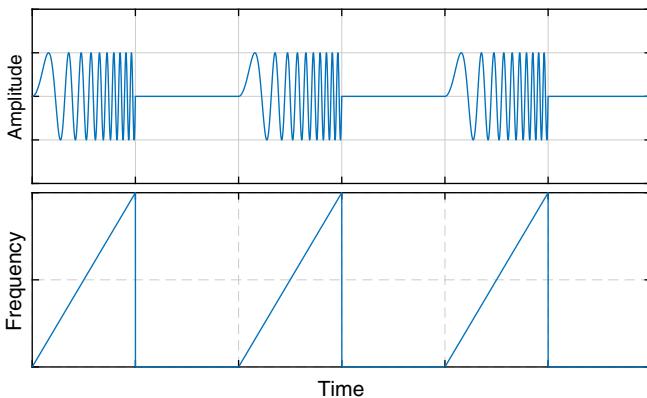
### 7.3.2 LINEAR FM AND FMCW RADAR

One way to increase the bandwidth of a pulse waveform is to rapidly sweep the carrier frequency within the pulse interval  $T_1$  [33]. In an LFM radar, the instantaneous frequency of a pulse has a linear dependency on time,

$$f(t) = \frac{B}{T_1} t, \quad 0 \leq t < \tau, \quad (7.8)$$

where  $T_1$  is the pulse duration and  $t$  is the time variable. Accordingly, the transmitted LFM radar signal is

$$\begin{aligned} s(t) &= \exp \left( 2\pi J f_o t + 2\pi J \int_0^t f(\hat{t}) d\hat{t} \right) \\ &= \exp \left( 2\pi J f_o t + \pi J \frac{B}{T_1} t^2 \right), \quad 0 \leq t < \tau, \end{aligned} \quad (7.9)$$

**FIG. 7.10**

An illustration of the amplitude and the frequency of an LFM burst of pulses.

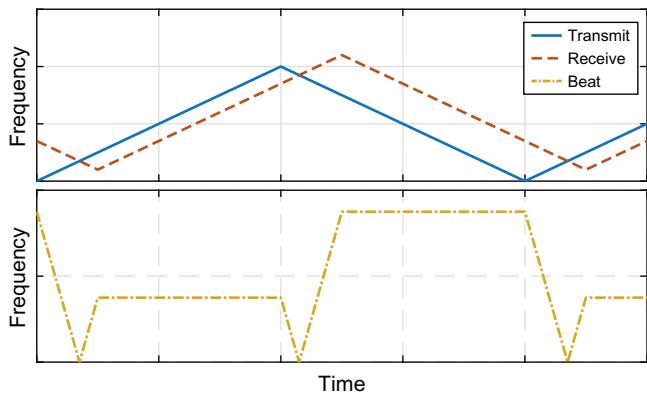
where  $\hat{t}$  is a dummy integration variable. LFM waveforms are often referred to as *chirps* because of their audible analog. A burst of LFM pulses is shown in Fig. 7.10, and the corresponding instantaneous frequency is shown in the lower part of the same figure. Because of this frequency sweep, an LFM waveform stretches over wider bandwidth than  $1/T_1$  and its time-bandwidth product is  $BT_1 \gg 1$ . However, pulse radar has low duty cycle because it transmits high power for very short intervals. One of the serious limitations of CMOS technology is the modest transmit power as we cannot apply high voltage to obtain high-power output [45]. Thus a continuous waveform is preferred, having a unity duty cycle, and a moderate bandwidth. Thus by stretching a moderate power output over an extended period, the scene can be illuminated with sufficient energy for a reliable detection. This approach allows the possibility of using CMOS technology for radar applications.

A popular continuous waveform is the frequency modulated continuous wave (FMCW) [46] that, unlike pulse waveforms, stretches the frequency sweep over an extended period of time. Similar to LFM, in FMCW the carrier frequency is linearly swept in some specific bandwidth  $B$ ; however, the time period  $T_1$  in FMCW is much larger than LFM and can be varied to have a specific variable slope  $q = B/T_1$ . Thus by controlling  $T_1$  we can get different frequency slopes.

FMCW can resolve both range and Doppler by mixing the received signal with a copy of the transmit signal. This technique can be easily implemented in low-cost CMOS chips. When the received and transmitted signals are mixed and low-pass filtered, a resulting low-frequency signal, also called the *beat signal*, is obtained having a frequency of

$$f_b = q \frac{2R}{c} + \frac{2\nu}{c} f_o, \quad (7.10)$$

where  $R$  and  $\nu$  are the range and velocity, respectively, of the target;  $c$  is the speed of light; and  $f_o$  is the center frequency (carrier frequency). The beat signal has a low



**FIG. 7.11**

An illustration of the frequency modulated continuous wave and the resulting beat signal. For nonstationary target the rising and declining parts produce different beat frequencies as per Eq. (7.10).

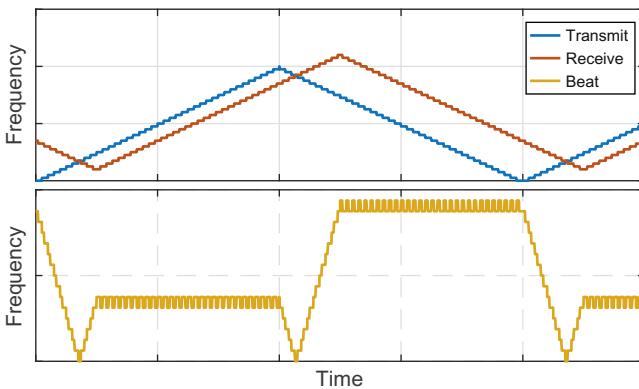
frequency in an order of megahertz, for automotive application. We depict in Fig. 7.11 the waveform of a typical FMCW signal with two cycles, one has a rising slope  $\varrho_1$ , and the next has a declining slope  $\varrho_2 = -\varrho_1$ . The figure also depicts the reflected signal and the beat signal resulting from mixing the transmitted signal with the reflected one. Having two different slopes facilitate the decoupling of range and Doppler information in Eq. (7.10). While the detection of multiple targets requires the use of multiple different slopes.

Implementing a linear voltage-controlled-oscillator VCO in CMOS chips imposes a considerable challenge because of some nonlinear chip components such as the varactor diode [45]. Nonlinearity can be overcome by utilizing a lookup table for the voltage or by using advanced signal processing techniques [47, 48]. However, the stability of the nonlinearity curve is itself an issue due to its dependency on temperature and load variations [45]. This unstable nonlinearity is different than the frequency unsuitability, which can be treated using an external reference crystal oscillator [45].

It is important to note that a stepped version of the FMCW radar arises when using a lookup table, due to the finite length of the table. This causes some undesirable ripples in the beat frequency, where a longer lookup table indeed leads to a smoother response. Fig. 7.12 shows a quantized version of the FMCW, note the effect of quantization on the beat signal.

### 7.3.3 STEPPED FREQUENCY RADAR

The principle of this method is to capture the scene in the *frequency domain* by sampling its transfer function at the frequencies of some transmitted tones  $F_n$ , then converting these samples using the inverse discrete Fourier transform (IDFT) into the

**FIG. 7.12**

An illustration of the stepped version of the frequency modulated continuous wave. Note the effect of quantization on the beat signal.

time domain [49]. Thus we obtain the temporal impulse response of the scene and detect the targets accordingly. This method has long been used in high-resolution radars (HRRs) [43] to counter the limitations in designing a single ultra-wideband pulse waveform by alternatively sending a frequency-stepped pulse train where each pulse has a narrow bandwidth.

In stepped frequency radar, a waveform is composed of multiple *tones*, or *waveform segments*, that are synthesized from a common stable oscillator and thus coherently related. Each of these transmitted tones lasts for a time duration  $T_1$  that should be long enough for the receiver to capture all reflected signals from the scene, starting from the closest reflection up to the farthest reflection. In other words, the radar scene is kept excited for a duration  $T_1$  to allow the interaction of the EM signals reflected from all targets. For example in automotive radar, if the scene extends for 150 m then  $T_1 > \frac{2 \times 150}{c} \approx 1 \mu\text{s}$ . However, in a time-varying scenario, a waveform needs to be properly designed to avoid *range-walk* problems [50] that occur due to the nonnegligible motion of targets.

The sequence of the transmitted frequencies in each tone can be coded in any desired manner. It is common to linearly increase the frequency from one tone to the next. In this case, the resulting waveform will look similar to the stepped-FMCW. However, the main difference is that the step duration in HRR should be long enough to properly sample the steady-state response of the scene, while in the stepped-FMCW it needs to be as short as possible in order for the mixer to produce a valid beat signal.

Instead of linearly increasing the frequency from a tone to the next one, a randomized step is often used to enhance the covert and reduce the interference problems [17]. This approach is called RSF and is achieved by randomly selecting the next tone in the pulse train from a pool of allowable tones.

### 7.3.4 PSEUDO-RANDOM STEPPED FREQUENCY RADAR

The essence of the PRSF radar is that the transmitted waveforms consist of a train of short tones with frequencies defined according to a pseudo-random sequence. This approach is very close to the concept of RFS. At one level, RSF and PRSF appear to be quite similar; however, there are important differences related to the type of receive processing and management of multiradar interference. PRSF sequences are easier to learn by other radars, thus enabling a radar to avoid the spectral collision. For RSF radar it is not possible to predict upcoming tones due to its purely random nature. In the PRSF case, it is feasible to exploit the structure to design sequences which are learnable, have minimal cross-correlation, and efficiently utilize the allocated spectrum.

Thus a PRSF radar selects a frequency sequence  $Q$  from a set of orthogonal sequences  $\mathcal{Q}$  that are permutations of the ordered set  $Q_o$  which is given by,

$$Q_o = \{f_o + \delta_f, f_o + 2\delta_f, \dots, f_o + N\delta_f\}, \quad (7.11)$$

where  $\delta_f = \frac{B}{N}$  is the tones spacing,  $B$  is the allocated bandwidth,  $f_o$  is the base frequency, and  $N$  is total number of tones. After choosing a desired (optimal) sequence, a radar transmits these tones sequentially, with a duration  $T_1$  of each tone.

Since a single tone has a constant frequency during the tone duration  $T_1$ , we can approximate the instantaneous bandwidth of the PRSF radar as the inverse of  $T_1$  (i.e.,  $B_1 \approx \frac{1}{T_1}$ ). In order to provide better immunity against radar-to-radar interference, we can select the tone spacing as  $\delta_f \geq B_1$ , so that a radar transmitting at a particular tone will have minimal interference to the adjacent tones of other radars. The transmitted signal in the interval  $(n - 1)T_1 < t \leq nT_1$  can be written as,

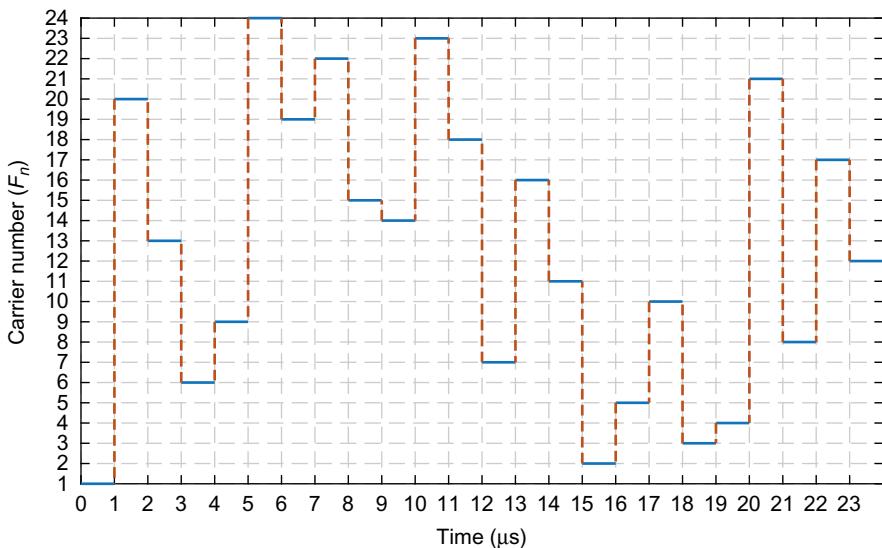
$$s_n(t) = \exp(j2\pi F_n t), \quad (7.12)$$

where  $F_n$  is the  $n$ th element in a selected sequence  $Q \in \mathcal{Q}$  from the set of orthogonal sequences. We depict in Fig. 7.13 an example of a random sequence spanning 24 carriers over 24  $\mu\text{s}$ . Fig. 7.14 depicts a time-domain snapshot of a PRSF waveform for a small number of tones.

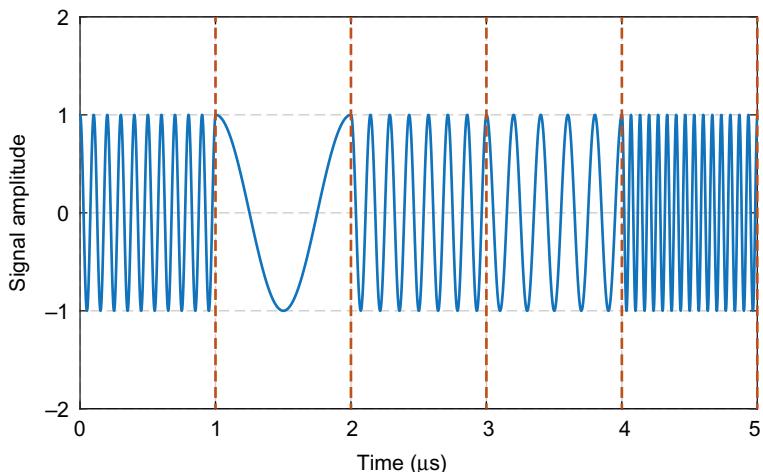
One of the requirements for the PRSF-based radar is that the returned PRSF signal should not interfere with the transmitting signal by other radars. This requirement can be met by using PRSF tone sequences with minimized Hamming cross-correlation. In case of a periodic PRSF of a period  $L$ , the Hamming cross-correlation function  $H_{AB}(k)$  between two sequences  $A = \{a_0, a_1, \dots, a_{L-1}\}$  and  $B = \{b_0, b_1, \dots, b_{L-1}\}$  is given by

$$H_{AB}(k) = \sum_{i=0}^{L-1} h\{a_i, b_{(i+k) \bmod L}\}; \quad k = 0, 1, \dots, L-1, \quad (7.13)$$

where  $k$  is a parameter representing the shift between the two sequences,  $h$  is a binary Hamming function given by  $h(a, b) = 1$  if  $a = b$  and  $h(a, b) = 0$  if  $a \neq b$ .

**FIG. 7.13**

Pseudo-random stepped frequency waveform.

**FIG. 7.14**

An illustration of the time domain of the pseudo-random stepped frequency waveform for five time intervals  $T_1$ .

The Hamming correlation between any two sequences  $A, B$  is as given in Eq. (7.13). We have the following two parameters concerning Hamming correlations:

$$\begin{aligned} \text{HA}_{\max} &= \max_{A \in Q} \{H_{AA}(\tau), \tau \neq 0\}, \\ \text{HC}_{\max} &= \max_{A, B \in Q} \{H_{AB}(\tau), \text{either } A \neq B \text{ or } \tau \neq 0\}. \end{aligned} \quad (7.14)$$

Ideally, we would like  $\text{HC}_{\max}$  to be zero, so that signals due to two or more radars do not interfere at any given point in time. However, there is a theoretical lower bound on  $\text{HC}_{\max}$ , first studied by Lempel and Greenberger [51]. The bound depends on specific arrangements of the sequence elements. Let  $A$  be a sequence over  $Q$  of a period  $L$  and let  $W(A) = [W_i(A), 1 \leq i \leq |Q|]$  be the weight vector associated with  $A$ , where  $W_i(A)$  is the number of  $i$ th symbol of  $Q$  appearing in  $A$ . Furthermore, let  $A$  and  $B$  be two sequences of period  $L$  over an alphabet  $Q$  of size  $|Q|$ , and let  $b$  be the least nonnegative residue of  $L \bmod |Q|$ . Then, for a family of two sequences  $\{A, B\}$

$$\text{HA}_{\max} \geq \frac{(L-b)(L+b-|Q|)}{|Q|(L-1)}, \quad (7.15)$$

$$\text{HC}_{\max} \geq \frac{\beta - 2L}{3L - 2}, \quad (7.16)$$

where  $\beta$  is related to the structure of the weight vectors of  $A$  and  $B$ , given by

$$\beta = \sum_{i=1}^{|Q|} \left[ W_i(A)W_i(B) + (W_i(A))^2 + (W_i(B))^2 \right]. \quad (7.17)$$

Note that we may assume  $W_1(A) \geq W_2(A) \geq \dots \geq W_{|Q|}(A)$ . Furthermore, the right-hand side of Eq. (7.16) is minimized whenever the following conditions are satisfied

- $W_1(A) - W_{|Q|}(A) \leq 1$ .
- $W_i(B)$ 's are in an increasing order  $W_1(B) \leq W_2(B) \leq \dots \leq W_{|Q|}(B)$ .
- $W_{|Q|}(B) - W_1(B) \leq 1$ .

The above bound at specific parameters yields interesting observations. When  $L = p^{mr} - 1$  and  $|Q| = p^{m\rho}$ , where  $p$  is a prime numbers;  $r, m$  are integers; and  $1 < \rho < r$ . The above inequalities simplify to [51],

$$\begin{aligned} \text{HA}_{\max} &\geq p^{m(r-\rho)} - 1, \\ \text{HC}_{\max} &\geq p^{m(r-\rho)}. \end{aligned} \quad (7.18)$$

It can be shown that when the size of the alphabet is almost equal to the period, it is possible to have zero autocorrelation. However, when the number of users in the system is two or more, the cross-correlation parameter has to be greater than or equal to 1. The specific case of the 1 coincidence sequences is used in practice extensively [52]. When the number of sequences in a family is greater than 2, Sarwate [53] derived the improved bound given following:

$$\text{HC}_{\max} \geq \frac{N}{|Q|} - \frac{1}{M}. \quad (7.19)$$

A recent paper by Peng and Fan [54] has a formulation similar to that of [53] and the bound is given as follows:

$$\text{HC}_{\max} \geq \frac{(LM - |\mathcal{Q}|)L}{(LM - 1)|\mathcal{Q}|}. \quad (7.20)$$

The above results apply when two PRSF radars are operating in the same spatial environment.

The reflected signal from point targets (scatterers) will consist of  $K$  echoes of the transmitted signal, with each echo  $k \in [1, K]$  is characterized by a complex amplitude  $\beta_k$  that captures both the magnitude change and the phase shift caused by the scatterer. Each echo has a delay  $\tau_k = \frac{2R_k}{c}$ , and a Doppler frequency shift  $u_k = \frac{2u_k}{c}f_0$ , where  $R_k$  and  $u_k$  are the range and velocity of the scatterer, respectively. Thus we write the received signal in the interval  $(n - 1)T_1 < t \leq nT_1$  as,

$$r_n(t) = \sum_{k=1}^K \beta_k s_n(t - \tau_k) \exp(j2\pi u_k t), \quad (7.21)$$

where  $s_n(t)$  is the transmitted PRSF signal described in Eq. (7.12).

### 7.3.5 PROCESSING A PRSF WAVEFORM

Since we know the transmitted frequency sequence, we can mix the received signal with a copy of the originally transmitted waveform, accordingly the resulting baseband signal in the interval  $(n - 1)T_1 < t \leq nT_1$  is as follows,

$$\begin{aligned} y_n(t) &= r_n(t)s_n^*(t) \\ &= \sum_{k=1}^K \beta_k \underbrace{\exp(-j2\pi F_n \tau_k)}_{\text{Range}} \underbrace{\exp(j2\pi u_k t)}_{\text{Doppler}}, \end{aligned} \quad (7.22)$$

noting the two components of the received signal: (i) a range component has a negligible change over the coherent processing interval (CPI) and (ii) an oscillating component because of the Doppler frequency having an order of few tens of kilohertz for a typical millimeter-wave automotive radar scenario. These time-dependent signals can be grouped in a vector  $\mathbf{Y}(t) = \{y_n(t)\}$ , where <sup>1</sup> $n = 1, \dots, N$ , as it will be used later in Eq. (7.23).

We assume that the number of detectable echoes is  $\kappa \leq K$ , where the aim of the signal processing is to estimate the range and Doppler these detectable echoes; that is, to find a vector of delays  $\psi = \{\tau_1, \dots, \tau_K\}$  and a corresponding vector of Doppler

---

<sup>1</sup>In this section, we use the *bold* notation to denote sets and vectors.

shifts  $\chi = \{u_1, \dots, u_K\}$  that minimizes the square discrepancies between the received signal and the estimated signal. Mathematically, these vectors are

$$\{\psi, \chi\} = \arg \min_{\psi, \chi} \left\| Y(t) - \left\{ \sum_{l=1}^K \exp(-j2\pi F_n \tau_l) \exp(j2\pi u_l t) \right\} \right\|_2, \quad (7.23)$$

implicitly including the estimation of the number of detectable targets  $K$ , which is equal to the length of the vectors  $\psi$  and  $\chi$ . The notation,

$$\|X\|_2 = \sqrt{|x_1|^2 + |x_2|^2 + \dots + |x_N|^2}, \quad (7.24)$$

represents the Euclidean norm of the complex vector  $X$ , and  $|x|^2 = xx^*$  is the square magnitude. Practical approaches to estimate the range and Doppler are further elaborated in Sections 7.3.5.1 and 7.3.6.

### 7.3.5.1 Waveform repetition for M-times

One practical method to extract the range and Doppler requires the repetition the PRSF waveform described in Eq. (7.12) for  $M$ -times within the CPI  $T_{\text{CPI}}$ , such that  $T_{\text{CPI}} = M \times T_2 = N \times M \times T_1$ , where  $T_2$  is the time required for  $N$  tones to be emitted (a single waveform). Thus the received signal of the  $m \in [1, M]$  waveform in a CPI during the time interval  $(n-1)T_1 < t \leq nT_1$  can be written as,

$$y_{mn}(t) = \sum_{k=1}^K \beta_k \exp(-j2\pi F_n \tau_k) \exp(j2\pi u_k(t + (m-1)T_2)). \quad (7.25)$$

A radar receiver detects the return phase and magnitude of each tone in every waveform, where the samples are organized in a complex matrix  $Y_{M \times N}$ , having  $N$  columns representing the tones and  $M$  rows representing the waveforms,

$$Y_{M \times N} = \left\{ \sum_{k=1}^K \beta_k \exp(-j2\pi F_n \tau_k) \exp(j2\pi u_k(m-1)T_2) \right\}_{m,n}, \quad (7.26)$$

where  $m \in [1, M]$  and  $n \in [1, N]$  refer to the rows and columns of the matrix, respectively. In order to obtain the ranges of targets we apply IDFT in a *row-wise* manner to transform the scene response from the frequency domain to the delay domain.

A processing-effective IDFT can be achieved by using the inverse *fast* Fourier transform (IFFT) algorithm. For this algorithm to work, the columns of the complex matrix  $Y_{M \times N}$  need to be rearranged corresponding to an ascending tone-frequency order. Then IFFT is applied in a row-wise manner (row-by-row) to the matrix  $Y_{M \times N}$ . Thus the result is

$$Z_{M \times N} = \text{ifft}[Y_{M \times N}, 2], \quad (7.27)$$

where  $\text{ifft}[\cdot, 2]$  means row-wise. The peaks in  $Z_{M \times N}$  occur at the delays  $\tau_k$ , thus the range of a target  $k$  is obtained as  $R_k = \frac{c\tau_k}{2}$ .

Another method can be applied which does not require the rearrangement of columns. In this method a *reference* matrix is correlated to the received matrix

$Y_{M \times N}$ , this reference matrix is a time-shifted version of the transmitted signal, having the form,

$$X_{I \times N} = \{\exp(-j2\pi F_n \tau_i)\}_{i,n}, \quad (7.28)$$

where  $i \in [1, I]$ ,  $n \in [1, N]$  refers to the rows and columns of the matrix, respectively, while the reference delays  $\tau_i$  are generated such that they sufficiently cover the non-ambiguous range of the radar, that is, if the target range is  $R \in [R_1, R_2]$  then the reference delay is calculated as,

$$\tau = \left[ 2\frac{R_1}{c}, 2\frac{R_1 + \Delta_R}{c}, 2\frac{R_1 + 2\Delta_R}{c}, \dots, 2\frac{R_2}{c} \right], \quad (7.29)$$

where  $\Delta_R$  is the desired range cell size.

The computation of the correlation between the received matrix and the reference matrix is done as follows,

$$\begin{aligned} Z_{M,I} &= Y_{M \times N} \times X_{I \times N}^\dagger \\ &= \left\{ \sum_{n=1}^N \sum_{k=1}^K \beta_k \exp[-j2\pi F_n(\tau_k - \tau_i)] \exp[j2\pi u_k(m-1)T_2] \right\}_{m,i}, \end{aligned} \quad (7.30)$$

where  $\dagger$  refers to the conjugate transpose of a matrix. Similar to Eq. (7.27), the peaks in  $Z_{M,I}$  will correspond to the targets in the scene. Indeed, this is apparent as when  $\tau_k = \tau_i$  the correlation will peak as the summation at these particular columns will have the form:

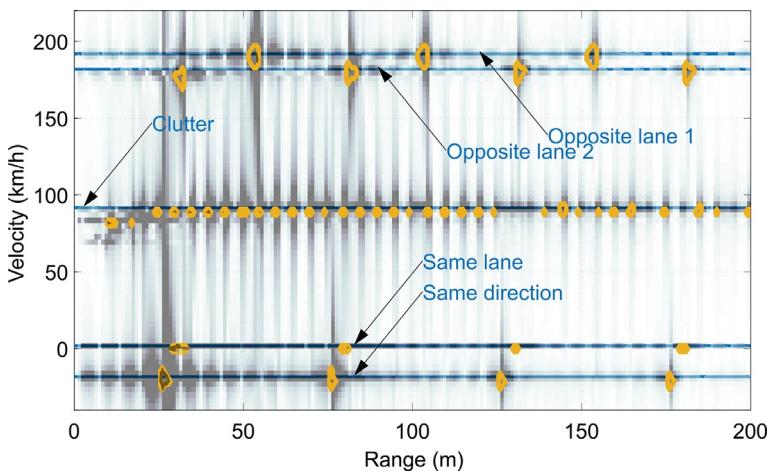
$$Z_{M,I}(\tau_k = \tau_i) = \left\{ N \sum_{k=1}^K \beta_k \exp(j2\pi u_k(m-1)T_2) \right\}_{m,i}. \quad (7.31)$$

A similar method is also presented in [17] by Axelsson describing the correlation approach to obtain the range information. However, Axelsson does not elaborate on how to obtain the Doppler information. It is worthy to indicate that this method is identical to performing the discrete Fourier transform (DFT) on  $Y_{M \times N}$ ; however, the transformation matrix should be rearranged in the same order of the tones.

Using any of the earlier two methods for extracting the range information leads to a two-dimensional matrix denoted as  $Z_{M,I}$  where each row represents the temporal response of the channel, while columns represent the slow change occurring to the temporal response because of the Doppler component. This can be seen from Eq. (7.31), where the peaks oscillate, within a column, according to the target Doppler frequency. In order to extract these oscillations, we apply a fast Fourier transform in a *column-wise* manner (i.e., column-by-column) and the result is another two-dimensional matrix containing the range-Doppler information,

$$\rho_{M \times I} = \text{fft}[Z_{M \times I}, 1], \quad (7.32)$$

where  $\text{fft}[, 1]$  means *column-wise*. An illustration of the range-Doppler matrix  $\rho_{M \times I}$  is depicted in Fig. 7.15, showing a simulation of a typical vehicular scenario, where targets have different Doppler frequencies depending on the relative velocities with respect to the frame of reference associated with the radar. Indeed in a practical



**FIG. 7.15**

An illustration of the range-Doppler matrix for a typical vehicular radar scenario, where a road composed of four lanes (two on each side) is simulated. Peaks are found using a peak detection algorithm indicating targets which are vehicles and the clutter on each side of the road.

receiver a noise component will affect the received signal and hence produce a noisy range-Doppler matrix. Also, Doppler ambiguities are generated as an effect of the waveform repetition time. In fact the nonambiguous Doppler interval can be calculated as the inverse of waveform repetition time,

$$\Delta u = \frac{1}{T_1}, \quad (7.33)$$

Eq. (7.33) indicates that a trade-off exists between waveform length and Doppler ambiguity.

It is important to note that in a time-varying scenario, the repetition of the waveform for  $M$ -times might create serious issues due to significant motion of targets within the total time  $T_{\text{CPI}} = M \times T_2$ . Thus adding additional requirement to compensate for targets motion [50]. An alternative approach is elaborated in [Section 7.3.6](#), where two different waveforms are emitted, the first is to detect significant Doppler shifts in the scene; and the second is to discover the targets associated with these Doppler shifts.

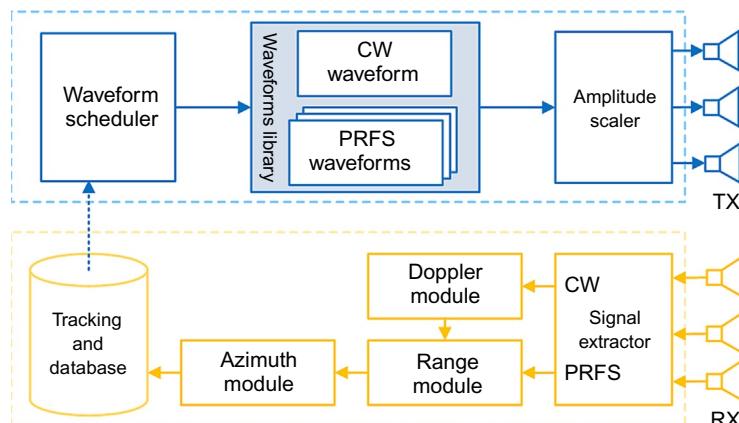
### 7.3.6 ADAPTIVE RADAR AND COMPUTATIONALLY LIGHT PROCESSING TECHNIQUES

Many applications of CMOS consumer radar require the identification of multiple objects over a wide field of view within a relatively short period of time, especially when considering highly dynamic applications such as automotive radars. These

requirements create many design challenges when finding the optimum balance between cost and performance. Advanced signal processing platforms for radars are indeed capable of handling complex real-time algorithms; however, both cost and form-factor constraints limit the usability of such platforms in consumer radars. Accordingly, it is essential to devise computationally light processing techniques that balance high performance and computational cost, in particular, such techniques should be designed to:

- detect targets with sufficient resolution in range, Doppler, and azimuth;
- reduce computational complexity requirements; and
- reduce interference between radars sharing the same spectrum.

Thus a general architecture of the main logical units of an adaptive radar is shown in Fig. 7.16 based on [18], where the system can employ multiple antennas to extract the azimuth information. Single antenna radars can still find applications in rear-facing vehicle parking system, anticollision system in UAVs and bicycles. The transmitter part of the design (upper part of Fig. 7.16) shows a waveform scheduler that first transmits a simple continuous wave (CW) to probe the scene for all detectable Doppler shifts, after which it selects from a set of available, predesigned, waveforms that which best suits the particular scene [55]. This method has two advantages: (i) the correlation test (or Fourier transform) needs to be run on few specific Doppler frequencies rather than the entire Doppler span; and (ii) the scheduler will only select the best suitable waveform rather than undertaking a full waveform design in real-time. Thus a significant reduction in the processing requirements can be achieved. The CW waveform and the PRFS waveform can be multiplexed (scheduled) either in the time domain (sequentially) or in the frequency domain and then mixed with the carrier signal having a frequency of  $f_0$ . An amplitude scaling is added in the front end to effectively increase the dynamic range. This scaling reduces the amplitude during



**FIG. 7.16**

Generic block diagram for logical components of adaptive consumer radar.

the sampling period of close targets, so that the receiver does not saturate and hence swamp the echoes from more distant targets.

The lower part of Fig. 7.16 shows the receiver section of the radar, where the reflected signal is collected by several antennas or a single antenna, then amplified by an LNA and mixed with the carrier signal  $f_o$  and further mixed with a baseband waveform related to the originally transmitted waveform. The final mixed signal is passed to the Doppler, range, and azimuth extraction modules, and the associated target information is stored and tracked in a targets database. The next scheduled waveform will utilize the dynamic targets database to select the most suitable waveform that detects the predicted locations of targets. The rest of the analysis in this section is mostly based on the patent [18].

### 7.3.6.1 Detection of significant Doppler frequencies

Consider the CW waveform emitted by the radar, having the form:

$$s(t) = A \exp(j2\pi f_o t), \quad (7.34)$$

where  $f_o$  is the carrier frequency. Accordingly, the reflected signal received by the radar is given by:

$$y(t) = \sum_{k=1}^K \beta_k s(t - \tau_k) \exp(j2\pi u_k t) + \varpi_1(t), \quad (7.35)$$

where similar to Eq. (7.21) in Section 7.3.5, several echoes  $K$  of the transmit signal are received corresponding to scatters  $k \in [1, K]$ . In this equation we treat the nonideal case when the received signal is impaired by Additive White Gaussian Noise (AWGN)  $\varpi_1(t)$ . It should be clarified that in a real-world scenario, clutter is present that produces a colored-like noise with characteristics that are not necessarily well described by Gaussian distributions. Nevertheless, simple clutter cancellation techniques may be implemented that are based on the knowledge of the platform speed. Relatively narrow antenna beam patterns and the forward looking geometry would limit the clutter spread and effective clutter filtering could be implemented at very little computational cost. After clutter cancellation, the assumption of AWGN becomes appropriate and the signal model in Eq. (7.35) can be considered accurate for the description of the received signal in nonideal conditions. The rest of the symbols is similar to Eq. (7.21). The CW signal extractor will mix the received signal with the carrier and perform sampling with a regular interval  $T_s$ . Accordingly, the resulting sampled vector is

$$\begin{aligned} z_1[n] &= y(nT_s) \exp(-j2\pi f_o nT_s) \\ &= \sum_{k=1}^K \beta_k \exp(-j2\pi f_o \tau_k) \exp(j2\pi u_k nT_s) + \varpi_1(nT_s) \\ &= \sum_{k=1}^K \beta'_k \exp(j2\pi u_k nT_s) + \varpi_1(nT_s). \end{aligned} \quad (7.36)$$

The samples of the AWGN are assumed independent with uniformly distributed phase angle  $\sim \mathcal{U}(-\pi, \pi)$ . The last step is achieved by including all time-invariant

constants in  $\beta'_k = \beta_k \exp(-j2\pi\tau_k) \in \mathbb{C}$ , noting that the scene is approximated as time-invariant within only a short period of time where the delays  $\tau_k$  are considered constants.

Thus the vector  $\mathbf{z}_1$  is used to obtain a rough estimation of the number of targets  $K$  and their Doppler frequencies  $u_k$ , which will substantially reduce the complexity of processing the subsequent PRSF waveform. In particular, the aim of the Doppler extraction module is to obtain the most significant Doppler frequency bins. Detecting the significant Doppler frequencies is thus achieved by iteratively comparing the normalized DFT of the received sequence  $z[n]$  with a predefined threshold  $\Gamma$ . Where the normalized DFT is calculated as,

$$\hat{\mathbf{Z}}_1[m] = \frac{1}{\bar{z}} \frac{|Z_1[m]|}{N}, \quad m \in [1, 2, \dots, N], \quad (7.37)$$

where  $Z_1$  is the discrete time Fourier transform of  $\mathbf{z}_1$ , that is,

$$Z_1[m] = \sum_{n=1}^N z_1[n] \exp\left(-j2\pi m \frac{n}{N}\right), \quad m \in [1, 2, \dots, N], \quad (7.38)$$

and  $\bar{z}_1$  is the mean magnitude of the time vector  $\mathbf{z}_1[n]$ ,

$$\bar{z}_1 = \frac{\sqrt{\mathbf{z}_1 \cdot \mathbf{z}_1^\dagger}}{N}, \quad (7.39)$$

where  $\dagger$  refers to the conjugate transpose. After obtaining the normalized FFT  $\hat{\mathbf{Z}}[m]$ , we search for the index of the highest Doppler component in that vector,

$$m_i^* = \arg \max_m [\hat{\mathbf{Z}}_1[m]]. \quad (7.40)$$

If the maximum frequency component is above a certain design threshold  $\Gamma$  then we subtract this component from the original time domain vector and update the vector such that,

$$\mathbf{z}_1[n] \leftarrow \mathbf{z}_1[n] - \hat{\mathbf{Z}}_1[m_i^*] \exp\left(j2\pi m_i^* \frac{n}{N}\right). \quad (7.41)$$

The detection algorithm should keep repeating these steps iteratively until the largest frequency component is below the threshold  $\Gamma$ . Thus the vector of all detectable digital Doppler frequencies is  $\mathbf{M}^* = \{m_1^*, \dots, m_i^*, \dots, m_K^*\}$ , where  $\kappa \leq K$  is the number of detectable targets. The approach of iteratively subtracting Doppler components is similar in principle to the successive interference cancellation method, where a receiver knows the structure of the interference [56].

### 7.3.6.2 Robust range-Doppler estimation

Once the significant Doppler frequencies have been estimated from the CW waveform, the scheduler transmits a PRSF waveform. It is worthy to mention that the number of Doppler frequencies provided by the Doppler module is not necessary the same as the number of targets, since several targets (scatters) could have a similar velocity. However, the range module should be able to resolve these targets subject to its range resolution.

The received and mixed signal from a PRSF waveform has the form described in Eq. (7.22), in addition to some AWGN noise  $\varpi_2(t)$ ,

$$y_n(t) = \sum_{k=1}^K \beta_k \exp(-j2\pi F_n \tau_k) \exp(j2\pi u_k t) + \varpi_2(t). \quad (7.42)$$

If the signal is sampled at intervals  $T_1$  we get,

$$z_2[n] = \sum_{k=1}^K \beta_k \exp(-j2\pi F_n \tau_k) \exp(j2\pi u_k n T_1) + \varpi_2(n T_1). \quad (7.43)$$

A correlation is applied to the sampled signal such that,

$$\rho(i, l) = \frac{1}{\bar{z}_2} \frac{|\mathbf{Z}_2[i, l]|}{N}, \quad i \in [1, 2, \dots, K], \\ l \in [1, 2, \dots, N], \quad (7.44)$$

where  $\mathbf{Z}_2$  is calculated from  $z_2$ , as,

$$\mathbf{Z}_2[i, l] = \sum_{n=1}^N z_2[n] \exp\left[-j2\pi \left(m_i^* \frac{n}{N} - F_n \frac{l}{N}\right)\right], \quad i \in [1, 2, \dots, K] \\ l \in [1, 2, \dots, N]. \quad (7.45)$$

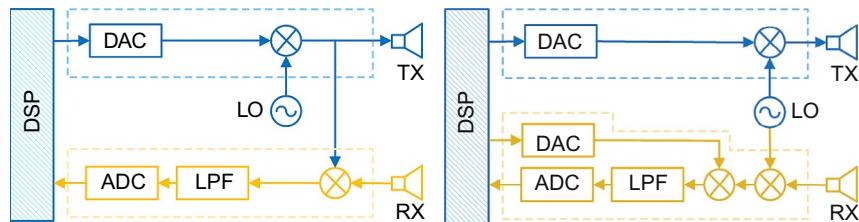
The peaks in the matrix  $\rho(i, l)$  will correspond to the range and Doppler of the targets. If the targets are not well-separated in the range-Doppler domain, then an iterative method needs to be applied as in the one applied for Doppler in Eq. (7.41).

The processing of a PRSF waveform with known Doppler significantly reduces the requirements of processing-power in CMOS chip, since a two-dimensional DFT is a computationally demanding operation. Readers are referred to the patent [18] for further details.

### 7.3.7 INTERMEDIATE FREQUENCY PROCESSING TECHNIQUE

The implementation technology of consumer radar on a single small CMOS chip imposes hard constraints on many of the design aspects. In addition to the limited signal processing capability comes the transmit power, and the phase noise of the local oscillator. On the other hand, the fulfillment of wide field of view and high angular resolution requires the exploitation of multiple receiver channels and a complex RF distribution network. Frequency modulated waveforms are therefore commonly employed to minimize the instantaneous transmit power and to simplify signal processing techniques.

As an alternative to FMCW we have explored the random step frequency (RSF) waveform in Section 7.3.3 and its pseudo-random derivative (PRSF) in Section 7.3.5 that mitigate the mutual interference when a group of radars share the same spectrum. Building an FMCW radar can be done using the simplified diagram in the left part of Fig. 7.17, where a digital-to-analog converter (DAC) generates a baseband tone that is steadily sweeping over the desired bandwidth. Multiple sweeps are required to detect targets on different range-Doppler locations, these sweeps are usually designed with different ramp slopes as discussed previously.

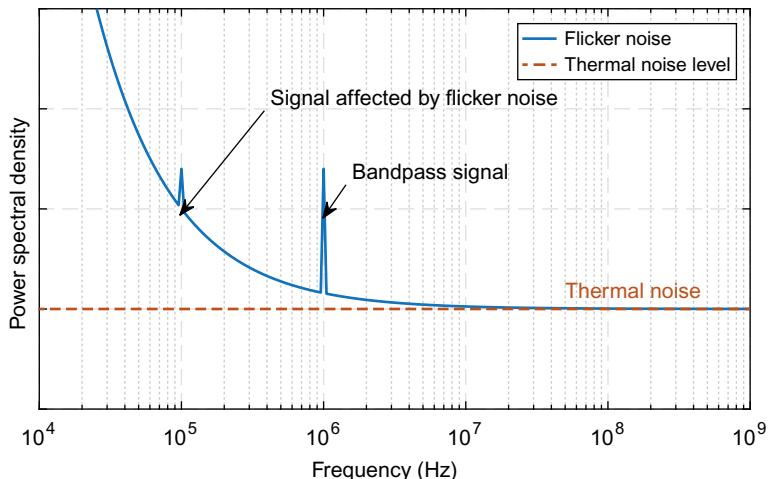
**FIG. 7.17**

Conventional methods for designing software defined radar system: (*left*) one-step down conversion; (*right*) two-step down conversion.

The diagram in the right part of Fig. 7.17 enhances the RF distribution performance by tapping from the local oscillator directly; however, this comes at the cost of having an additional DAC that reproduces the transmitted signal, which will have an impact on the chip size and cost. The baseband signal in the left design will be affected by *flicker noise* when the frequencies of interest are very close to DC. Flicker noise, also known as the  $1/f$ -noise, is defined to have a spectral power density of the following form for near the DC region [57]:

$$N(f) = \frac{\text{Constant}}{f^\alpha}, \quad (7.46)$$

where  $\alpha \approx 1$ , and  $f$  is the frequency. In practice, the flicker noise converges to Gaussian noise for higher frequency as the thermal noise effect exhibits *white* spectrum density  $N_0$ . The plot in Fig. 7.18 shows a signal at low frequency significantly

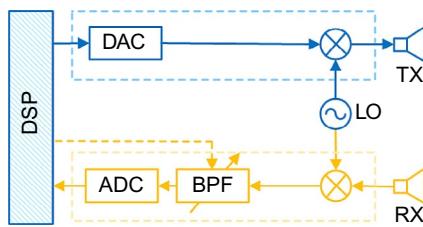
**FIG. 7.18**

An illustration of the flicker noise and its effect on low-frequency signals.

affected by flicker noise, having a low SNR, and another signal at higher frequency with better SNR performance.

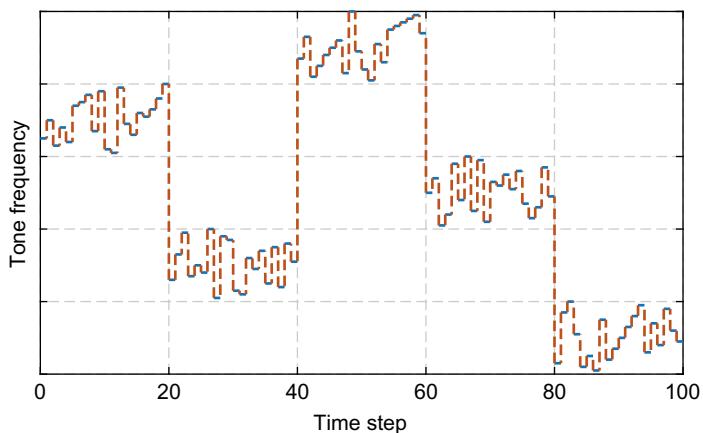
In order to overcome this deficiency, the design in the right part of Fig. 7.17 can be utilized with the DAC generating a fine-resolution intermediate frequency (IF) mixing signal. This secondary mixing frequency should be chosen such that the resulting signal is down-converted very close to the baseband with a proper separation from the  $1/f$ -noise. However, as stated previously, the DAC solution requires additional space on the chip and will affect the cost.

An alternative design is suggested in Fig. 7.19 based on the patent in [19] which employs a switched, or tunable, band-pass filter and performs the sampling in an IF band slightly above the base band. This design can effectively overcome the flicker noise problem with reduced chip design complexity [19]. The utilized tunable band-pass filter comprises of a bank of fixed frequency band-pass filters and a mechanism to switch between them. Taking the PRSF as an example, the waveform needs to be divided into several subgroups as indicated in Fig. 7.20, where the constraints on the



**FIG. 7.19**

Improved platform for CMOS radar chip, with a tunable, or switchable, band-pass filter.



**FIG. 7.20**

Random stepped frequency radar waveform with band-pass grouping.

speed of switching the center frequency and the available tuning range limit the frequency sequence that may be used. Ideally, the filter center frequency should be switchable for every pulse in the sequence of the signal. If the filter does not respond sufficiently fast for this, the sequence is limited to using tones in each filter band in some sequence before jumping to another band, potentially placing some constraints on the degree of randomness of the PRSF signal that can be employed.

---

## 7.4 STOCHASTIC GEOMETRY TECHNIQUE FOR MODELING AUTOMOTIVE CONSUMER RADARS

Taking automotive radar as an important future application of consumer CMOS radars, we present in this section some techniques and methods to estimate the potential interference and we show how to optimize system parameters for best radar performance using stochastic geometry.

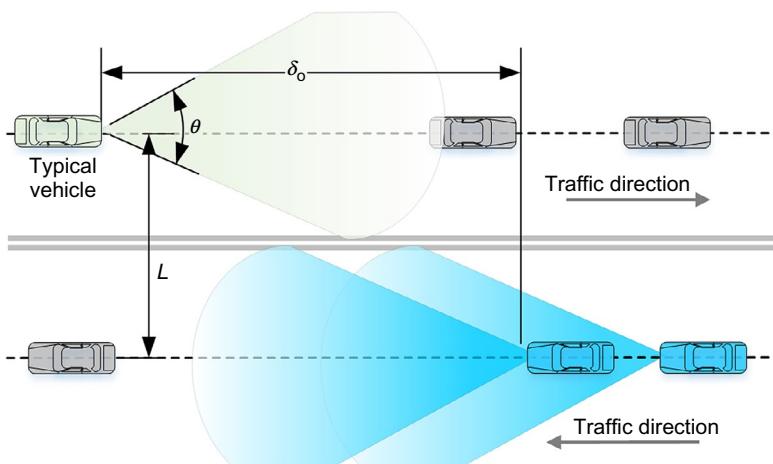
Stochastic geometry deals with random points and gives great insight into different parameters that affect the overall performance of the system. Vehicles traveling on a road can also be seen as random points, where using tools from stochastic geometry can capture the arising interference in terms of its cumulative distribution function (CDF) and mean value based on a given vehicle density in a road segment. This approach can be further employed to understand the average performance of automotive radar in terms of *ranging success probability*, that is the probability of reliably detecting a target given a certain set of operating conditions. In this section we show some explicit formulae that tightly characterize the lower bound performance and provide insight into the different dynamics influencing performance. Furthermore we show an optimizing technique to find the optimum duty cycle for any specific radar to randomly access spectrum resources.

A simplified layout of the interfering long-range radars is shown in Fig. 7.21 showing a typical vehicle with the potential interfering radars traveling in the opposite direction. Taking into consideration the defined narrow antenna pattern and ignoring sidelobes, the interfering vehicles are those located beyond a minimum distance  $\delta_0$ , expressed as:

$$\delta_0 = \frac{L_n}{\tan \frac{\theta}{2}}, \quad (7.47)$$

where  $\theta$  is the antenna beamwidth,  $L_n$  is the distance between the lane of the typical vehicle and the  $n$ th opposing lane, where multiple opposing lanes can exist.

We can capture the randomness in the locations of vehicles in two extreme geometrical distributions (point processes). In the first case we assume complete irregularity in the locations of vehicles with no correlation between these locations. This case is modeled by a Poisson point process (PPP) with intensity  $\lambda$ . The second extreme occurs when vehicles are located on a deterministic lattice layout, that is,

**FIG. 7.21**

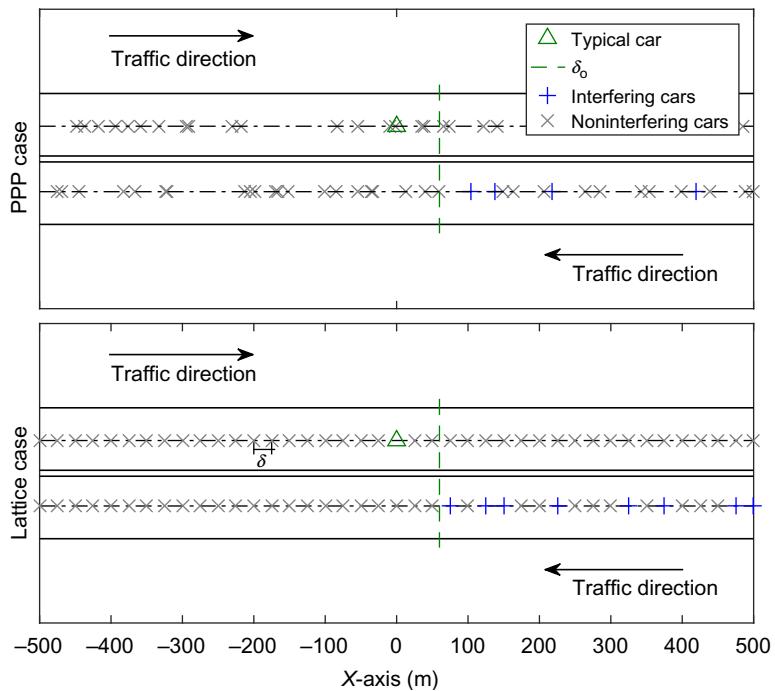
A simplified geometrical layout indicating the typical vehicle under study and interfering vehicles.

periodic locations in space, separated by a constant distance  $\delta$ . In this layout vehicle locations are considered to be completely regular. In practical scenarios we expect that the actual distribution of vehicles will reside between these two geometrical extremes. Based on the preceding model, we depict a simplified geometrical layout in Fig. 7.22 showing both point processes and indicating the typical vehicle location and interfering vehicle locations. We now consider the two extreme point processes further.

### 7.4.1 POISSON POINT PROCESS MODEL

One extreme of the geometrical distribution is achieved when vehicle locations on a certain lane are completely independent of each other. This case resembles a unidimensional PPP in  $\mathbb{R}^1$  with a homogeneous linear intensity  $\lambda$  measured in vehicles per unit length. We denote this set of vehicles as  $\Phi_{\text{PPP}}$ . Utilizing a PPP allows a tractable analysis to be developed using Campbell theorems [58]. To model the effect of medium access, thus to capture the effect of concurrently transmitting vehicles, we apply *random thinning* on the opposing vehicles set  $\Phi_{\text{PPP}}$  with a retention probability equal to the probability of resource collision given as  $\xi$ , representing the potential that an opposing vehicle is concurrently utilizing the same resources as the typical vehicle, thus  $\xi$  can be thought of as the duty cycle of the random spectrum access. Applying a random marking on  $\Phi_{\text{PPP}}$  we can describe the interferers set as:

$$\Theta_{\text{PPP}} = \{x: x \in \Phi_{\text{PPP}}, \mathcal{M}(x) = 1\}, \quad (7.48)$$

**FIG. 7.22**

The proposed geometrical models: (*up*) a linear Poisson point process and (*down*) a regular lattice.

where the mark  $\mathcal{M}(x)$  is defined as:

$$\mathcal{M}(x) = \begin{cases} 0: & x \leq \delta_0 \\ \mathbf{B}(\xi): & x > \delta_0, \end{cases} \quad (7.49)$$

where vehicles closer than  $\delta_0$  are marked as noninterfering, and  $\mathbf{B}(\xi)$  is a Bernoulli random variable (RV) with selection probability  $\xi$ , where RVs in this chapter are denoted in *bold*.

### 7.4.2 LATTICE MODEL

In this model we assume that vehicles are distributed according to a deterministic (regular) one-dimensional lattice, where vehicles can only take discrete locations with predefined spacing distance  $\delta$  unit length, thus having a linear density of  $\lambda = \frac{1}{\delta}$ . Although we assumed that vehicles' locations within a lane are deterministic, they would exhibit no correlation to the typical vehicle in the opposing lane. Thus we assume that a uniformly distributed RV randomly translates the entire lattice in a linear manner. Accordingly, we may express the set of approaching vehicles as:

$$\Phi_{TL} = \{(x + \mathcal{U})\delta + \delta_0 : x \in \mathbb{Z}\}, \quad (7.50)$$

where  $\mathbb{Z}$  is the set of integers, and  $\mathcal{U}$  is a standard uniformly distributed RV in the range  $[0, 1]$ . It should be noted that  $\mathcal{U}$  is single RV (not a vector) that captures the randomness in the grid translation with respect to the typical vehicle, where all approaching vehicles are translated with an equal value of  $\mathcal{U}\delta$ . Following Eq. (7.49) we can *mark* the subset of interfering vehicles:

$$\Theta_{BL} = \{x: x \in \Phi_{TL}, \mathcal{M}(x) = 1\}, \quad (7.51)$$

where the subscript BL means *Bernoulli lattice*. This is depicted in the lower part of Fig. 7.22, where interfering vehicles are indicated with a (+) sign.

### 7.4.3 INTERFERENCE ANALYSIS

In order to find the mean (expectation) value of the interference, we apply the Campbell theorem [58] to calculate the sum over a PPP:

$$\begin{aligned} \bar{I}_{PPP} = \mathbb{E}[I] &= \mathbb{E}_g \mathbb{E}_{\Theta_{PPP}} \left[ \sum_{x \in \Theta_{PPP}} g_x \gamma_1 P_o ||x||^{-2} \right] \\ &\stackrel{(a)}{=} \mathbb{E}_g[g] \int_{\delta_o}^{\infty} \lambda_1 \gamma_1 P_o u^{-2} dr \\ &= \int_{\delta_o}^{\infty} \lambda_1 \gamma_1 P_o u^{-2} dr, \end{aligned} \quad (7.52)$$

where  $\mathbb{E}_g$  is the expectation over the channel stochastic process,  $\mathbb{E}_{\Theta_{PPP}}$  is the geometric expectation over all possible realizations of the interferers' locations, and  $\lambda_1 = \xi \lambda$  is the effective intensity (density) of the interferers. The variable  $u = \sqrt{r^2 + L_n^2}$  represents the distance between the typical vehicle and the interferers in the  $n$ th opposing lane. Step (a) in Eq. (7.52) follows the assumption that individual propagation channels have an i.i.d. distribution, which is independent of the geometrical point process. The final step assumes that the average channel gain is normalized to unity, that is,  $\mathbb{E}[g] = 1$ . Evaluating the integral yields,

$$\bar{I}_{PPP} = \frac{\lambda \xi \gamma_1 P_o \left( \pi - 2 \tan^{-1} \left( \frac{\delta_o}{L_n} \right) \right)}{2L_n}. \quad (7.53)$$

If we neglect the lane spacing when compared to the longitudinal distance  $r$ , we can obtain,

$$\bar{I}_{PPP}|_{L_n \rightarrow 0} = \frac{\xi \lambda \gamma_1 P_o}{\delta_o}, \quad (7.54)$$

we note that the interference is linearly proportional to the effective interferer density  $\xi \lambda$ . Also we note the strong effect of  $\delta_o$ , where the interference is inversely proportional to  $\delta_o$ . This is clear that a narrower antenna beamwidth will increase  $\delta_o$  and reduce the interference.

Similar to the PPP case, we can obtain the interference arising from BL.<sup>2</sup> However, an interesting case appears when we ignore the lane distance when compared to

---

<sup>2</sup>Please refer to the detailed proof in [59].

the longitudinal span of the road (i.e.,  $L_n \rightarrow 0$ ), where it can be shown that the average interference from a translated Bernoulli lattice is exactly equal to its counterpart on the linear PPP.

#### 7.4.4 INTERFERENCE STATISTICS

In order to get a deeper insight into the statistical behavior of the interference, we can obtain its characteristic function (CF) and then formulate its CDF. Thus it can be shown that the CF of interference is given by Ref. [59],

$$\varphi_{I_{\text{PPP}}}(\omega)|_{\text{wc}} = \exp\left(-\sqrt{-j\pi\gamma_1 P_o \omega \lambda_i^2}\right), \quad (7.55)$$

having a tractable so-called Lévy distribution which can be seen as an inverse gamma distribution [60], having a CF and a CDF of the form:

$$\begin{aligned} \varphi(\omega) &= \exp\left(j\mu\omega - \sqrt{-2j\alpha\omega}\right), \\ F_X(x) &= \text{erfc}\left(\sqrt{\frac{a}{2(x-\mu)}}\right), \end{aligned} \quad (7.56)$$

where  $\mu$  is the location parameter,  $a$  is the scale parameter, and  $\text{erfc}(z)$  is the complementary error function. By comparing Eqs. (7.55), (7.56) we can conclude that the interference follows a Levy distribution with CDF:

$$F_I(x)|_{\text{wc}} = \text{erfc}\left(\sqrt{\frac{\pi(\xi\lambda)^2 \gamma_1 P_o}{4x}}\right), \quad (7.57)$$

and with location parameter  $\mu = 0$  and scaling parameter  $a = \frac{1}{2}\pi(\xi\lambda)^2 \gamma_1 P_o$ .

#### 7.4.5 PERFORMANCE ANALYSIS AND OPTIMIZATION

If the aggregated interference  $I$  summed at the receiver is uncorrelated with the transmit signal, then the receiver perceives this interference as noise-like, thus causing an increase in the noise floor. This behavior is analytically investigated for FMCW radars in [14] as well as emphasized by ITU-R in [21]. The performance of radar is thus limited by the signal to interference and noise ratio, defined as:

$$\text{SINR} = \frac{S}{I + N}, \quad (7.58)$$

where  $N$  is the noise power process generated in the receiver electronics, and  $S$  is the reflected power from the target having the form described previously in Section 7.2.2. Having the SINR above a certain threshold  $T$  leads to successful ranging and detection. Accordingly we form the probability of successful ranging as:

$$\begin{aligned} p_s &= \mathbb{P}[\text{SINR} \geq T] \\ &= \mathbb{P}\left[I \leq \frac{S}{T} - N\right] = F_I\left(\frac{S}{T} - N\right). \end{aligned} \quad (7.59)$$

Thus the probability of successful ranging takes the following closed form,

$$p_s|_{wc} = \operatorname{erfc} \left( \sqrt{\frac{\frac{\pi}{4}(\xi\lambda)^2 \gamma_1 P_o}{\frac{\gamma_1 \gamma_2 P_o R^{-4}}{T} - N}} \right). \quad (7.60)$$

For dense traffic conditions and sufficient radar power, the limiting performance factor becomes the interference rather than the noise. In this case, the ranging success probability will take a simpler form,

$$p_s|_{wc, N=0} = \operatorname{erfc} \left( \sqrt{\frac{\pi T}{4\gamma_2} \xi \lambda R^2} \right). \quad (7.61)$$

This provides an insight on the main dynamics affecting the performance of automotive radar. In an interference-limited environment, the ranging success probability is independent of the common transmit power  $P_o$  and the antenna characteristics  $\gamma_1$ . We plot in Fig. 7.23 the ranging probability as a function of the ranging distance  $R$ , and the vehicles' intensity  $\lambda$  using the closed form in Eq. (7.61) and compared with Monte-Carlo simulations.

As it can be deduced from Eq. (7.61) a higher spectrum collision probability  $\xi$  leads naturally to a lower radar success rate. However by reducing  $\xi$ , the spectrum utilization efficiency will proportionally drop, since fewer vehicles are concurrently accessing the spectrum. If vehicles utilize a random spectrum access policy then  $\xi$  can also be seen as the transmission probability over the shared bandwidth. Finding

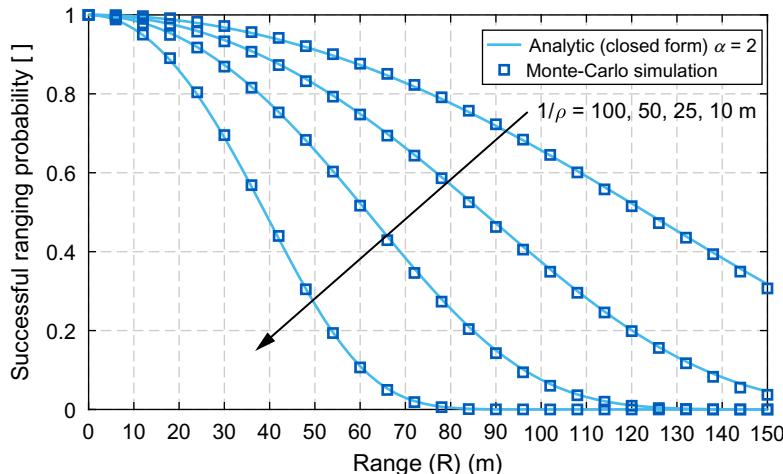


FIG. 7.23

The ranging success probability  $p_s$  in Eq. (7.61) versus the ranging distance for the worst case scenario, using  $\xi = \frac{1}{100}$ .

an optimum design value of  $\xi$  can substantially enhance the overall system performance for all users.

We define the *spatial success probability*  $\beta$  as the probability of successful spectrum access per unit length expressed as,

$$\beta = \lambda \xi p_s, \quad (7.62)$$

representing the density of vehicles that are detecting their targets successfully. Having units of *success per unit length*. The *success* in this context is defined as *successful access of the spectrum*. Recalling that  $\lambda_I = \lambda \xi$ , we plot in Fig. 7.24 the spatial success probability against  $\lambda_I$ , and the target range  $R$ . It should be noted that a certain optimum density point exists for a certain target range, where operating at this point leads to maximizing the spatial success probability.

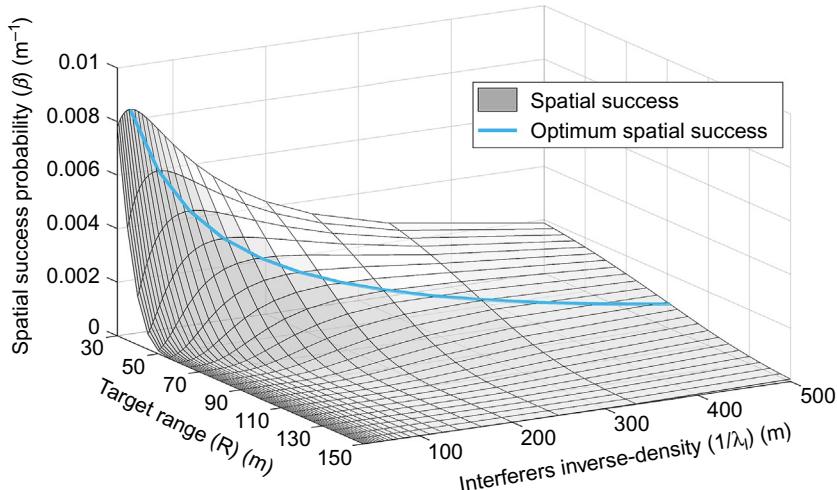
As  $\beta$  is a concave function, we formulate the optimum intensity as,

$$\lambda_I^* = \operatorname{argmax}_{\lambda_I} [\beta = \lambda_I \operatorname{erfc}(C\lambda_I)], \quad (7.63)$$

taking  $\frac{\partial \beta}{\partial \lambda_I} = 0$  yields,

$$\operatorname{erfc}(C\lambda_I^*) = \frac{2C\lambda_I^* e^{-C^2 \lambda_I^{*2}}}{\sqrt{\pi}}, \quad (7.64)$$

where  $C = \sqrt{\frac{\pi T}{4\gamma_2}} R^2$ . However, no exact closed form solution for Eq. (7.64) exists, alternatively we define a new variable  $z_o = C\lambda_I^*$  and substitute in Eq. (7.64) leading



**FIG. 7.24**

The spatial success probability  $\beta$  given in Eq. (7.62) versus the ranging distance  $R$  and the interferers density  $\lambda_I$ , showing the corresponding optima.

to a numerical solution of  $z_0 \approx 0.532$ , accordingly the optimum transmission probability is:

$$\xi^* = \min \left[ \frac{z_0}{\lambda C}, 1 \right] = \min \left[ \frac{z_0}{\lambda} \sqrt{\frac{4\gamma_2}{\pi T}} R^{-2}, 1 \right], \quad (7.65)$$

where the function  $\min [.,.]$  ensures that the transmission probability is less than unity. Accordingly, the optimum spatial success probability is given by substituting in Eq. (7.62),

$$\begin{aligned} \beta^* &= \lambda \xi^* \operatorname{erfc}(C \lambda \xi^*) \\ &= \lambda \min \left[ \frac{z_0}{\lambda C}, 1 \right] \operatorname{erfc} \left( \lambda C \min \left[ \frac{z_0}{\lambda C}, 1 \right] \right). \end{aligned} \quad (7.66)$$

However, the ranging distance  $R$  is a stochastic quantity with statistics depending on the vehicles' linear density  $\lambda$ , where the  $n$ th nearest vehicle has a known closed-form distribution in a PPP process of,

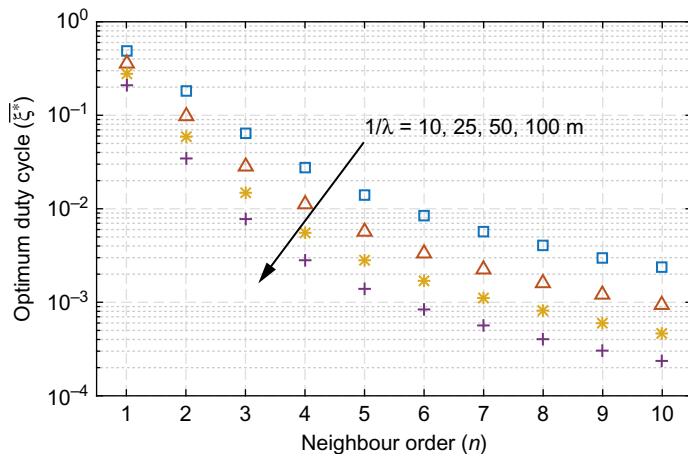
$$f_{R_n}(r) = \frac{e^{-\lambda r} (\lambda r)^n}{r \Gamma(n)}, \quad \forall n \in \mathbb{N}^+, \quad (7.67)$$

that follows from Eq. (2.21) in Ref. [60], where  $n \in \mathbb{N}^+$  represents the order of the nearest vehicle within the same lane. We can optimist the transmission probability (or duty cycle) by setting a certain target number of nearest vehicles to be detected, so the average optimum is obtained by performing a statistical expectation over the contact distance  $R$ ,

$$\begin{aligned} \bar{\xi}^* &= \mathbb{E}_R[\xi^*] \\ &= \int_0^\infty \min \left[ \frac{z_0}{\lambda} \sqrt{\frac{4\gamma_2}{\pi T}} R^{-2}, 1 \right] \times \frac{e^{-\lambda r} (\lambda r)^n}{r \Gamma(n)} dr \\ &= \frac{\lambda K \Gamma(n-2, \sqrt{\lambda K}) - \Gamma(n, \sqrt{\lambda K}) + \Gamma(n)}{\Gamma(n)}, \end{aligned} \quad (7.68)$$

where  $K$  is a constant given by  $K = z_0 \sqrt{\frac{4\gamma_2}{\pi T}}$ , and  $\Gamma(a, x) = \int_x^\infty t^{a-1} e^{-t} dt$  is the incomplete gamma function. Using different vehicle densities, we plot in Fig. 7.25 the optimum transmission probability (or duty cycle) at which the maximum spatial ranging success rate is achieved. As expected, the optimum transmission duty cycle drops with measured detection range. This becomes more clear if we study the asymptotic function of  $\bar{\xi}^*[n]$  in Eq. (7.68) as  $n \rightarrow \infty$ ,

$$\begin{aligned} \hat{\xi}[n] &= O(\bar{\xi}^*[n]) \quad \text{as } n \rightarrow \infty \\ &\stackrel{(a)}{=} O_{n \rightarrow \infty} \left( 1 + \frac{\lambda K e^{-\sqrt{\lambda K}} e_{n-2}(\sqrt{\lambda K})}{(n-1)(n-2)} - e^{\sqrt{\lambda K}} e_n(-\sqrt{\lambda K}) \right) \\ &= \frac{\lambda K}{n^2}, \end{aligned} \quad (7.69)$$

**FIG. 7.25**

Optimum duty cycle  $\xi^*$  given in Eq. (7.68) versus the design value of the target  $n$ th neighbor, for different vehicle densities.

step (a) follows from  $\Gamma(n, x) = (n - 1)!e^{-x}e_n(x)$  for  $n \in \mathbb{N}^+$ , and  $e_n(x) = 1 + x + x^2/2! + \dots + x^n/n!$  is the partial sum of the exponential series. Thus we can clearly note from Eq. (7.69) that detecting farther vehicles (higher  $n$ ) will decrease the spectral efficiency. However, it is interesting to note that an increased vehicle density leads to a better spectrum utilization, also apparent in Fig. 7.25 comparing the  $\xi$  for different values of  $\lambda$ .

## 7.5 PERFORMANCE LIMITATIONS

### 7.5.1 CMOS TECHNOLOGY LIMITATIONS

The capabilities of a transceiver built using CMOS technology are modest in terms of transmit power, oscillator phase noise, LNA dynamic range, and noise figure. Nevertheless, with innovative waveform design and sophisticated use of signal processing, it appears possible to achieve a high performance short-range single chip radar system capable of detecting relatively small targets and estimating their range, Doppler, and azimuth with reasonable accuracy. In this section we discuss the technological limitations on designing the best possible LNA having the lowest noise figure  $F$  and the highest transmit power  $P_t$ . For a given CMOS transistor channel length  $L$ <sup>3</sup>, carrier mobility  $\mu$  and a saturation electric field strength  $E_{\text{sat}}$ ,<sup>4</sup> we can write the expression of its cut-off angular frequency as follows [61, Eq. (5.41)],

<sup>3</sup>Channel length is the distance between the transistor's drain and source related to the photolithography resolution of the bulk CMOS substrate.

<sup>4</sup> $E_{\text{sat}}$  is defined as the electric field strength at which the carrier velocity drops below half of its value at low-field mobility, measured in V/m.

$$\omega_T \approx \frac{3\mu E_{\text{sat}}}{4L} = \frac{K_t}{L}, \quad (7.70)$$

where  $K_t = \frac{3}{4}\mu E_{\text{sat}}$  is a technology-related constant. The noise figure of CMOS LNA can be expressed as [61, Eq. (12.22)],

$$F \approx 1 + \frac{2}{\sqrt{5}\omega_T} \sqrt{\gamma\delta(1 - c_1)} = 1 + K_F \frac{\omega}{\omega_T}, \quad (7.71)$$

where  $\gamma$  is a dimensionless coefficient related to the drain current noise in a short channel in the ideal case,  $\delta$  is a coefficient related to the gate current noise, and  $c_1$  is the correlation coefficient between noise current in the drain and the gate. Thus  $K_F$  is another technology-related constant. The transmitter power is proportional to the square of the channel length,

$$P_t = K_\delta L^2 = K_\delta K_t^2 \frac{1}{\omega_T^2}. \quad (7.72)$$

The gain of an antenna is proportional to its effective aperture area  $A_e$ . The gain is given by,

$$G = \frac{4\pi A_e}{\lambda^2} = \frac{1}{\pi c^2} A_e \omega^2, \quad (7.73)$$

where  $\lambda$  is the wavelength and  $c$  is the propagation speed of the EM signal (speed of light). Thus the SNR of a radar can be obtained using the modified radar equation for millimeter-wave using Eq. (7.1) as follows,

$$\text{SNR} = \underbrace{\frac{1}{\pi^3 (4c)^2}}_{\text{Constant}} \underbrace{\frac{K_\delta K_t^2 A_e^2}{k_b T B \left(1 + K_F \frac{\omega}{\omega_T}\right)}}_{\text{Technology/frequency}} \underbrace{\left(\frac{\omega}{\omega_T}\right)^2}_{\text{Target/absorption}} \underbrace{\frac{\sigma_c}{R^4} \eta(R)}_{\text{Target/absorption}}, \quad (7.74)$$

where  $k_b$  is Boltzmann constant,  $\sigma_c$  is the RCS area,  $R$  is the target distance, and  $B$  is the waveform bandwidth. For further details, readers are referred to [62]. This equation provides an insight on the different dynamics related to the technology that affects CMOS radar-on-a-chip performance, namely:

- CMOS technology affecting the cut-off frequency  $\omega_T$ .
- The operating frequency  $\omega$ .
- The chip size affecting antenna area  $A_e$ .

### 7.5.2 INFORMATION THEORY LIMITATIONS

After Shannon had proved the possibility of constructing codes that allow the transmission of messages in the presence of noise without error up to limits set by the available bandwidth, the transmitted power, and the SNR, there was an increase in interest in the connection of these mathematical ideas to the underlying physical processes. The information carrying capacity of the EM field subject to the limits set

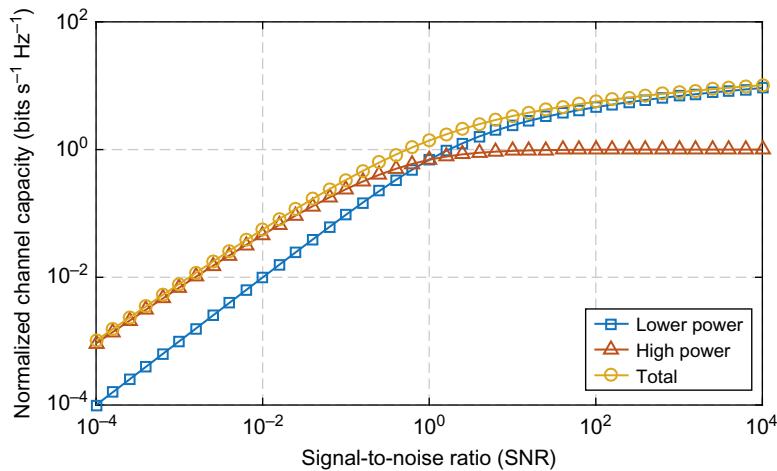
by quantum mechanics and statistical mechanics was found by constructing states of the field which maximized the physical entropy. By these means, Lebedev and Levitin [63] showed the channel capacity,  $C_p$ , of an EM field which was not constrained by bandwidth was given by

$$C_p = \frac{\pi}{\ln(2)} \sqrt{\frac{2P}{3h}} \quad (7.75)$$

and in the case of a bandwidth limited field by

$$C_p = B \log_2 \left( 1 + \frac{P}{hf_0 B} \right) + \frac{P}{hf_0} \log_2 \left( 1 + \frac{hf_0 B}{P} \right), \quad (7.76)$$

where  $P$  is the average power received by the detector,  $f_0$  is the carrier frequency of the bandwidth-limited signal,  $B$  is the bandwidth, and  $h$  is Planck's constant. The optimum spectrum for the bandwidth unlimited case is that of a black body radiator with a temperature set by the average power. This gives an absolute upper limit to the channel capacity given an available power. The bandwidth-limited case gives more insight into the limits for circumstances that are routinely used in radar and communications. The limit contains two terms. The first dominates in the limit of high power,  $P > hf_0 B$ , and has the form of the classic version of Shannon's channel capacity with the noise  $hf_0 B$  being the vacuum fluctuation noise in the bandwidth  $B$ . The second term is scaled with the photon arrival rate  $\frac{P}{hf_0}$ , and this term dominates when  $P < hf_0 B$ . Fig. 7.26 illustrates the channel capacity contributions of the two terms. It is



**FIG. 7.26**

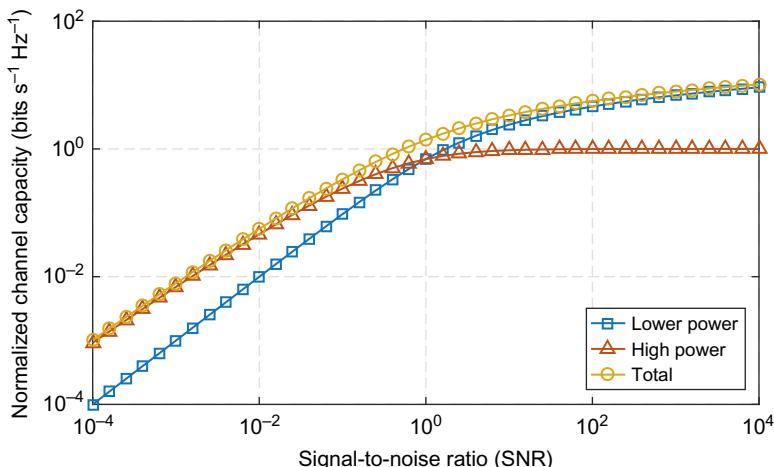
The normalized channel capacity  $\frac{C_p}{B}$  of the electromagnetic field in units of bandwidth as a function of the SNR  $\gamma = \frac{P}{hf_0 B}$  illustrating the high power  $P > hf_0 B$ , low power  $P < hf_0 B$ , and the total.

important that the only noise considered here is that set by the limits of quantum mechanics, the vacuum fluctuation noise. Any technical noise carried by the field is considered to be information about the source of that noise. In addition to this, the ability to transfer this information from the field into some other form is not guaranteed. The transfer of information from the field to an antenna and through an amplifier and filter to an analog to digital converter will place further limits on the information, which can be extracted from this field. The details of these limits are the subject of further study by the authors.

The ultimate limits of the radar system presented earlier can be calculated subject to these provisos and are shown in Fig. 7.27, which shows the channel capacity for an EM field versus received power with frequency 77 GHz and a bandwidth of 1 GHz and the channel capacity for the same power with no bandwidth limit. The power shown is the average which is incident on the receive antenna. In order to achieve such a channel capacity the transmitted EM field would have to be modulated in a manner such that after the interaction with whatever is being observed by the radar system, the received field was in this maximum entropy state. This would require either near perfect knowledge of the state of the scene under observation or a modulation scheme which adapts to the information which is obtained from continuing observations.

A simple estimate of the limits imposed by the transfer of power from the field to the antenna may be found by considering the channel capacity of the antenna subject to the EM noise in the environment. If we take the noise as due to the temperature of the background radiation in the bandwidth of the antenna, and the bandwidth of the antenna to be the same as that of the incident signal we have a channel capacity which may be written as

$$C_a = B \log_2 \left( 1 + \frac{P_a}{kTB} \right), \quad (7.77)$$



**FIG. 7.27**

The channel capacity for an electromagnetic field versus received power with frequency 77 GHz and a bandwidth of 1 GHz. Also shown is the bandwidth unlimited case.

where  $C_a$  is the channel capacity of the antenna,  $P_a$  is the power transferred from the field to the antenna,  $T$  is the temperature of the background radiation as seen by the antenna, and  $k$  is Boltzmann's constant. We are interested in the circumstance such that  $C_a = C_p$  subject to  $P_a \leq P$ . If all the radiation incident on the antenna is converted to electrical energy (i.e.,  $P_a = P$ ), we can find a condition for the maximum allowed temperature experienced by the antenna to achieve equality of channel capacity,  $T_{\max}$  given by

$$T_{\max} = \frac{hf_0}{k} \left( \frac{1}{\gamma+1} \right)^{\gamma} (\gamma+1) - 1, \quad (7.78)$$

with SNR  $\gamma = \frac{P}{hf_0 B}$  as defined previously. This maximum allowed temperature is shown for a 77 GHz, and 1 GHz bandwidth signal in Fig. 7.28 comparing with the cosmic microwave background radiation (CMBR) temperature [64], it is clear that the antenna needs to be very cold for this maximum power transfer to be feasible. It is also apparent that the required antenna temperature reaches an asymptotic upper bound given by

$$T_{\text{asym.}} = \lim_{\gamma \rightarrow \infty} T_{\max}(\gamma) = \frac{f_0 h}{e k}, \quad (7.79)$$

where  $e$  is the Euler's constant. The asymptotic maximum temperature is shown against the frequency in Fig. 7.29 illustrating the fact that for all powers at 77 GHz, the antenna temperature needs to be colder than the CMBR temperature. However, the antenna will also see the CMBR so there is an unavoidable additional

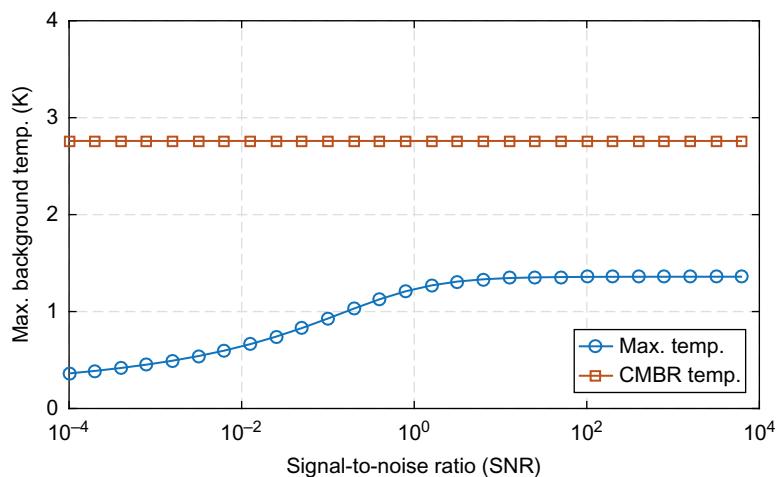
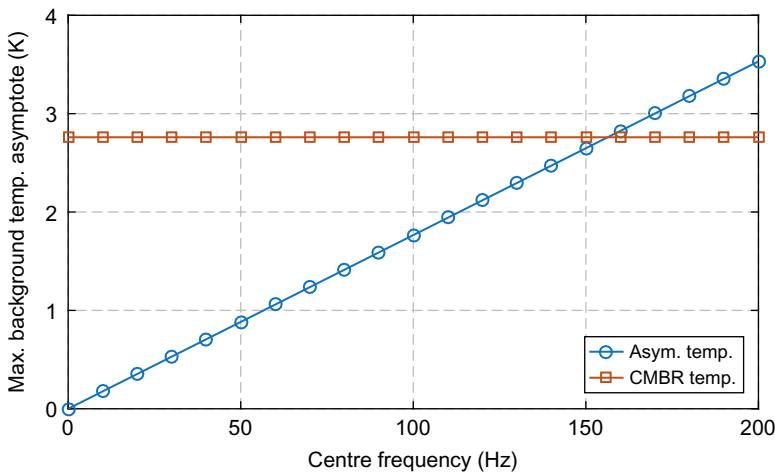


FIG. 7.28

The maximum background temperature  $T_{\max}$  observed by the antenna which allows complete matching of the field channel capacity to the antenna.

**FIG. 7.29**

The upper SNR asymptote of the maximum allowed antenna temperature as a function of carrier frequency.

noise over and above the quantum noise and the thermal noise in the antenna. Interestingly, we can obtain the frequency that is just high enough to allow the theoretical maximum power transfer when CMBR exists in the background. This frequency is given by,

$$f_0(T_{\text{CMBR}}) = \frac{ekT_{\text{CMBR}}}{h} \approx 156.32 \text{ GHz.} \quad (7.80)$$

Indeed, a practical radar antenna pointed toward a target experiences much higher background temperature than the CMBR; for example, the brightness temperature of the ground is taken around 240–290 K depending on the actual ambient temperature [65].

Thus it appears that the theoretical performance limit sets by quantum mechanics is much higher than the technological performance limits set by CMOS technology, the gap between the two bounds represents the feasible amount of performance enhancement that could be achieved if some optimum signal processing technique is utilized. Finding this optimum technique remains as the fundamental challenge of all times.

---

## ACKNOWLEDGMENTS

This research is funded by the Australian Research Council Discovery Project grant “Cognitive Radars for Automobiles,” ARC grant number DP150104473.

---

## REFERENCES

- [1] R.J. Evans, P.M. Farrell, G. Felic, H.T. Duong, H.V. Le, J. Li, M. Li, W. Moran, E. Skafidas, Consumer radar: opportunities and challenges, in: 11th European Radar Conference (EuRAD), 2014, pp. 5–8.
- [2] IEE Symposium on Consumer Applications of Radar and Sonar, 2013.
- [3] Project soli (Google). <https://atap.google.com/soli/>, 2016 (accessed 04.08.17).
- [4] M. Wicks, E. Mokole, S. Blunt, B. Schneible, Principles of Waveform Diversity and Design, SciTech, Raleigh, NC, USA, 2010.
- [5] M. Hartmann, C. Wagner, K. Seemann, J. Platz, H. Jäger, R. Weigel, A low-power low-noise single-chip receiver front-end for automotive radar at 77 GHz in silicon-germanium bipolar technology, in: 2007 IEEE Radio Frequency Integrated Circuits (RFIC) Symposium, IEEE, 2007, pp. 149–152.
- [6] J. Lee, Y.-A. Li, M.-H. Hung, S.-J. Huang, A fully-integrated 77-GHz FMCW radar transceiver in 65-Nm CMOS technology, *IEEE J. Solid State Circuits* 45 (12) (2010) 2746–2756.
- [7] V.H. Le, H.T. Duong, A.T. Huynh, C.M. Ta, F. Zhang, R.J. Evans, E. Skafidas, A CMOS 77-GHz receiver front-end for automotive radar, *IEEE Trans. Microw. Theory Tech.* 61 (10) (2013) 3783–3793.
- [8] B. Neri, S. Saponara, Advances in technologies, architectures, and applications of highly-integrated low-power radars, *IEEE Aerosp. Electron. Syst. Mag.* 27 (1) (2012) 25–36.
- [9] H.T. Duong, H.V. Le, A.T. Huynh, R.J. Evans, E. Skafidas, Design of a high gain power amplifier for 77 GHz radar automotive applications in 65-nm CMOS, in: Microwave Integrated Circuits Conference (EuMIC), 2013 European, 2013, pp. 65–68.
- [10] G.K. Felic, E. Skafidas, R. Evans, Metal plate lens antenna for automotive radar at Mm-wave frequencies, in: 2012 6th European Conference on Antennas and Propagation (EUCAP), IEEE, 2012, pp. 2321–2323.
- [11] R.J. Evans, P.M. Farrell, G. Felic, H.T. Duong, H.V. Le, J. Li, M. Li, W. Moran, M. Morelande, E. Skafidas, Consumer radar: technology and limitations, in: 2013 International Conference on Radar, ISSN 1097-5764, 2013, pp. 21–26.
- [12] M. Skolnik, Radar Handbook, third ed., McGraw-Hill, New York, NY, 2008.
- [13] M. Kunert, The EU Project MOSARIM: a general overview of project objectives and conducted work, in: 9th European Radar Conference (EuRAD), 2012, pp. 1–5.
- [14] G.M. Brooker, Mutual interference of millimeter-wave radar systems, *IEEE Trans. Electromagn. Compat.* 49 (1) (2007) 170–181.
- [15] T.-N. Luo, C.-H.E. Wu, Y.-J.E. Chen, A 77-GHz CMOS automotive radar transceiver with anti-interference function, *IEEE Trans. Circuits Syst. I Reg. Papers* 60 (12) (2013) 3247–3255.
- [16] T.-N. Luo, C.-H.E. Wu, Y.-J.E. Chen, A 77-GHz CMOS FMCW frequency synthesizer with reconfigurable chirps, *IEEE Trans. Microw. Theory Tech.* 61 (7) (2013) 2641–2647.
- [17] S.R.J. Axelsson, Analysis of random step frequency radar and comparison with experiments, *IEEE Trans. Geosci. Remote Sens.* 45 (4) (2007) 890–904.
- [18] M.R. Morelande, L. Mei, R.J. Evans, A method of target detection. CA Patent App. CA 2,865,803, Available from: <https://www.google.com/patents/CA2865803A1?cl=en>, 2013 (accessed 04.08.17).
- [19] R.J. Evans, J.Z.C. Li, E. Skafidas, W. Moran, M.R. Morelande, Apparatus and a method for obtaining information about at least one target. US Patent App. 14/357,638, Available from: <http://www.google.com/patents/US20140313068>, 2014 (accessed 04.08.17).

- [20] International Telecommunications Union (ITU), Characteristics of ultra-wideband technology. ITU-R SM.1755, Available from: <https://www.itu.int/rec/R-REC-SM.1755/en>, 2006 (accessed 04.08.17).
- [21] International Telecommunications Union (ITU), Systems characteristics of automotive radars operating in the frequency band 76–81 GHz for intelligent transport systems applications. ITU-R M.2057-0, Available from: <http://www.itu.int/rec/R-REC-M.2057-0-201402-I>, 2014 (accessed 04.08.17).
- [22] ACMA, Radiocommunications (low interference potential devices) class licence 2015, Tech. Rep., Australian Communications and Media Authority, 2015.
- [23] Australian Communications and Media Authority, A review of automotive radar systems—devices and regulatory frameworks, Tech. Rep., 2001.
- [24] ITU-R, Rec. ITU-R P.676-9 attenuation by atmospheric gases, P Series, Radiowave propagation, 2013.
- [25] Z. Qingling, J. Li, Rain attenuation in millimeter wave ranges, in: 2006 7th International Symposium on Antennas, Propagation EM Theory, 2006, pp. 1–4.
- [26] S.-C. Kim, B.J. Guarino, T.M. Willis, V. Erceg, S.J. Fortune, R.A. Valenzuela, L. W. Thomas, J. Ling, J.D. Moore, Radio propagation measurements and prediction using three-dimensional ray tracing in urban environments at 908 MHz and 1.9 GHz, IEEE Trans. Veh. Technol. 48 (3) (1999) 931–946.
- [27] A. Al-Hourani, S. Chandrasekharan, G. Baldini, S. Kandeepan, Propagation measurements in 5.8 GHz and pathloss study for CEN-DSRC, in: 2014 International Conference on Connected Vehicles and Expo (ICCVE), ISSN 2378-1289, 2014, pp. 1086–1091.
- [28] A.F. Molisch, F. Tufvesson, J. Karedal, C.F. Mecklenbrauker, A survey on vehicle-to-vehicle propagation channels, IEEE Wirel. Commun. 16 (6) (2009) 12–22.
- [29] M. Richards, J. Scheer, W. Holm, Principles of modern radar, SciTech Pub., Raleigh, NC, 2010.
- [30] V.N. Pozhidaev, Estimation of attenuation and backscattering of millimeter radio waves in meteorological formations, J. Commun. Technol. Electron. 55 (11) (2010) 1223–1230.
- [31] K.D. Ward, S. Watts, R.J.A. Tough, Sea clutter: scattering, the K distribution and radar performance, vol. 20, IET, London, UK, 2006.
- [32] J.B. Billingsley, A. Farina, F. Gini, M.V. Greco, L. Verrazzani, Statistical analyses of measured radar ground clutter data, IEEE Trans. Aerosp. Electron. Syst. 35 (2) (1999) 579–593.
- [33] M.A. Richards, Fundamentals of Radar Signal Processing, Tata McGraw-Hill Education, 2005.
- [34] M. Kronauge, H. Rohling, Fast two-dimensional CFAR procedure, IEEE Trans. Aerosp. Electron. Syst. 49 (3) (2013) 1817–1823.
- [35] A. Polychronopoulos, A. Amditis, N. Floudas, H. Lind, Integrated object and road border tracking using 77 GHz automotive radars, IEE Proc. Radar Sonar Navig. 151 (6) (2004) 375–381.
- [36] M.C. Budge, M.P. Burt, Range correlation effects in radars, in: Record of the 1993 IEEE National Radar Conference, 1993, pp. 212–216.
- [37] P.M. Woodward, Probability and Information Theory, With Applications to Radar, vol. 3, Pergamon Press LTD, London, UK, 1953.
- [38] C.H. Wilcox, The synthesis problem for radar ambiguity functions, Inst. Math. Its Appl. 32 (1991) 229.
- [39] S.M. Sussman, Least-square synthesis of radar ambiguity functions, IRE Trans. Inf. Theory 8 (3) (1962) 246–254.

- [40] D. Vakmann, Sophisticated Signals and the Uncertainty Principle in Radar, vol. 4, Springer Science & Business Media, New York, NY, 2012.
- [41] D. Cochran, S.D. Howard, B. Moran, Operator-theoretic modeling and waveform design for radar in the presence of Doppler, in: 2012 IEEE Radar Conference (RADAR), IEEE, 2012, pp. 0774–0777.
- [42] G. Turin, An introduction to matched filters, *IRE Trans. Inf. Theory* 6 (3) (1960) 311–329.
- [43] D.R. Wehner, High Resolution Radar, Artech House, Inc., Norwood, MA, 1987, 484 p.
- [44] D. Koks, How to create and manipulate radar range-Doppler plots, Tech. Rep., DTIC Document 2014
- [45] T. Mitomo, N. Ono, H. Hoshino, Y. Yoshihara, O. Watanabe, I. Seto, A 77 GHz 90 nm CMOS transceiver for FMCW radar applications, *IEEE J. Solid State Circuits* 45 (4) (2010) 928–937.
- [46] H. Rohling, C. Moller, Radar waveform for automotive radar systems and applications, in: 2008 IEEE Radar Conference, ISSN 1097-5659, 2008, pp. 1–4.
- [47] A. Suhre, T. Hammel, U. Luebbert, An adaptive method for compensating non-linear VCO characteristics using series reversion, in: 2015 16th International Radar Symposium (IRS), ISSN 2155-5745, 2015, pp. 155–160.
- [48] L. Li, J. Yu, J. Krolik, Software-defined calibration for FMCW phased-array radar, in: 2010 IEEE Radar Conference, ISSN 1097-5659, 2010, pp. 877–881.
- [49] C. Nguyen, J. Park, Stepped-Frequency Radar Sensors: Theory, Analysis and Design, Springer, Dordrecht, 2016.
- [50] Y. Liu, H. Meng, H. Zhang, X. Wang, Motion compensation of moving targets for high range resolution stepped-frequency radar, *Sensors* 8 (5) (2008) 3429–3437.
- [51] A. Lempel, H. Greenberger, Families of sequences with optimal hamming correlation properties, *IEEE Trans. Inf. Theory* 20 (1974) 90–94.
- [52] J.O.M. Simon, R. Scholtz, Spread Spectrum Communications Handbook, McGraw-Hill, New York, NY, 2002.
- [53] D.V. Sarwate, Optimum PN sequences for CDMA systems, in: IEEE Third International Symposium on Spread Spectrum Techniques and Applications, 1994, IEEE ISSSTA '94, vol. 1, 1994, pp. 27–35.
- [54] D. Peng, P. Fan, Lower bounds on the hamming auto- and cross correlations of frequency-hopping sequences, *IEEE Trans. Inf. Theory* 50 (9) (2004) 2149–2154.
- [55] D. Cochran, S. Suvorova, S.D. Howard, B. Moran, Waveform libraries, *IEEE Signal Process. Mag.* 26 (1) (2009) 12–21.
- [56] S.P. Weber, J.G. Andrews, X. Yang, G. de Veciana, Transmission capacity of wireless ad hoc networks with successive interference cancellation, *IEEE Trans. Inf. Theory* 53 (8) (2007) 2799–2814.
- [57] M.S. Keshner, 1/f noise, *Proc. IEEE* 70 (3) (1982) 212–218.
- [58] D. Stoyan, W.S. Kendall, J. Mecke, Stochastic Geometry and Its Applications, John Wiley, London, 1987.
- [59] A. Al-Hourani, R. Evans, S. Kandeepan, B. Moran, H. Eltom, Stochastic geometry methods for modelling automotive radar interference, *IEEE Trans. Intell. Transp. Syst.* (2016), <https://doi.org/10.13140/RG.2.1.4849.1128>.
- [60] M. Haenggi, Stochastic Geometry for Wireless Networks, Cambridge University Press, New York, NY, 2012, ISBN 1107014697.

- [61] T.H. Lee, *The Design of CMOS Radio-Frequency Integrated Circuits*, Cambridge University Press, New York, NY, 2004.
- [62] M. Li, R.J. Evans, E. Skafidas, B. Moran, Radar-on-a-Chip (RoaCH), in: 2010 IEEE Radar Conference, ISSN 1097-5659, 2010, pp. 1224–1228.
- [63] D.S. Lebedev, L.B. Levitin, Information transmission by electromagnetic field, *Inf. Control* 9 (1) (1966) 1–22.
- [64] D.J. Fixsen, The temperature of the cosmic microwave background, *Astrophys. J.* 707 (2) (2009) 916.
- [65] G. Maral, M. Bousquet, *Satellite Communications Systems: Systems, Techniques and Technology*, John Wiley & Sons, London, 2011.

# Signal processing for massive MIMO communications

# 8

**Muhammad R.A. Khandaker, Kai-Kit Wong**

*University College London, London, United Kingdom*

---

## 8.1 INTRODUCTION

Massive multiinput multioutput (MIMO), also known as large-scale antenna systems, is one of the exciting technologies enabling the fifth generation (5G) and beyond fifth generation (B5G) wireless cellular networks [3], which is limited, not by the number of antennas, but rather by the lack of sophisticated techniques to acquire channel-state information (CSI) for an unlimited number of terminals [1]. For next-generation wireless networks, it promises significant gains in terms of spectral efficiency (SE) and energy efficiency (EE) to accommodate a large number of users at extraordinarily high data rates with better reliability while consuming much less power. This chapter aims to give an overview of and address various important signal processing aspects of massive MIMO systems.

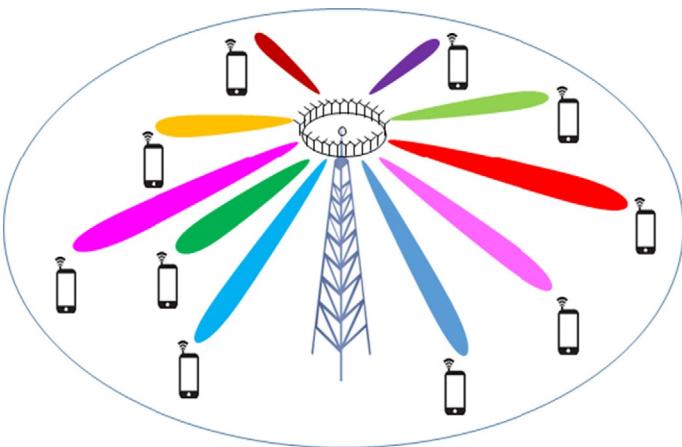
Multiple antenna systems are termed as massive MIMO systems when each base station (BS) is equipped with orders of magnitude more antennas than in systems being built today, often tens or hundreds in number. The BS with massive MIMO sends independent data streams to multiple user equipments (UEs) in the same time-frequency resource, hence achieving ultra-high SE. As the name suggests, very large MIMO involves an unprecedented number of antennas  $N$  simultaneously serving a much smaller number of terminals  $K$  (i.e.,  $N \gg K$ ) [2, 2, 3]. This exciting research area of future wireless communications has been sprouted by the seminal works in [4, 5] in an attempt to achieve more dramatic gains as well as to simplify the required signal processing.

Massive MIMO systems preferably operate in time-division duplexing (TDD) mode meaning that the uplink and downlink transmissions take place in the same frequency resource but are separated in time [6]. The physical propagation channels are often assumed to be reciprocal such that the channel responses are considered the same in both uplink (UL) and downlink (DL), which facilitates the TDD operation. Our emphasis on TDD rather than frequency-division duplexing (FDD) is inspired by the fact that we need to acquire CSI between an extreme number of service antennas and much smaller number of terminals. In addition, the time required to transmit

uplink pilots is independent of the number of transmit antennas  $N$ , whereas the time required to transmit downlink pilot is proportional to  $N$ . Hence in practice, massive MIMO systems exploit channel reciprocity to estimate the channel responses on the uplink and then use the acquired CSI for both uplink receive combining and downlink transmit precoding of actual data [6]. However, a practical limitation is that the transceiver hardware is generally not reciprocal. Hence calibration is necessary to exploit the channel reciprocity in practice. Fortunately, it has been observed in a 100-antenna testbed for massive MIMO that the uplink-downlink hardware mismatches only change by a few degrees over a 1-h period and hence can be mitigated by simple relative calibration methods [6, 7].

One of the many impressive features of massive MIMO is that as the number of BS antennas goes to infinity, the performance of linear precoders approaches to the performance of optimal nonlinear ones [5, 8, 9]. We will elaborate the mathematical foundation of this heuristic benefit later in this chapter. Following we summarize the basic advantages brought by massive MIMO from theory to practice [2, 3, 10].

- *Multiplexing gain:* Enormous spatial multiplexing provided by unlimited number of BS antennas supporting tens of users in massive MIMO systems makes it theoretically possible to increase the wireless capacity by 10 times or more [8]. A general rule of thumb is that massive MIMO systems should have an order of magnitude more antennas  $N$  than the number of scheduled users  $K$  because the users' channels are likely to be near-orthogonal when  $N/K > 10$ . Recently, an analysis on how the optimal number of scheduled users  $K$  depends on  $N$  has been performed in [11] and the value of optimal  $K$  has been derived in closed form.
- *SE:* Massive MIMO is a promising technique for increasing the SE of cellular networks. The large number of service antennas at the BS and multiplexing to tens of users rather than beamforming to a single user provides the SE gain [11, 12]. Even the use of moderately large antenna arrays can improve the SE with orders of magnitude compared to a single-antenna system [12]. Recently, an efficient system-level analysis has been performed in [11] to derive new SE expressions for multicell massive MIMO systems considering power control, arbitrary pilot reuse, and random user locations.
- *EE:* Huge energy saving is another inherent benefit of massive MIMO technology. The massive antenna arrays can potentially reduce UL and DL transmit powers through coherent combining and an increased antenna aperture [12]. It is already known that deploying multiantenna technique, UL transmit power of each UE can be reduced inversely proportional to the number of antennas at the BS with no adverse effect in performance [12]. It has been shown in [12] that each single-antenna user in a massive MIMO system can scale down its transmit power proportional to the number of antennas at the BS with perfect CSI or to the square root of the number of BS antennas with imperfect CSI, to get the same performance as a corresponding single-input single-output (SISO) system. However, massive BS antennas only reduce the radiation

**FIG. 8.1**

Beamforming in a typical massive MIMO system.

power, which only account for a very small portion as compared to the power consumption associated with the large number of radio frequency (RF) chains. Hence, RF chain reducing schemes has also been proposed [13].

- *Increased robustness and reliability:* The large number of antennas mounted at the BS, if proper precoding technique is applied, allows steering of directional beams toward the targeted UEs (as shown in Fig. 8.1). This in turn results in better performance in terms of received signal-to-noise ratio (SNR) and/or link reliability. Also, as we will investigate later, with infinitely many transmit antennas, the effect of uncorrelated noise, fast fading, and intracell interference disappears [2, 3, 5].
- *Simple linear processing:* When the number of BS antennas grows much greater than the number of UEs (i.e.,  $N \gg K$ ), simplest linear precoders and detectors approach optimal performance bounds [2, 5, 11, 14]. This saves significant processing resources, which would otherwise be used for precoder and/or decoder design and optimization.
- *Cost reduction in power components:* Since the energy consumption in massive MIMO systems is greatly reduced, the large array of antennas allows the use of very low-cost RF amplifiers in the milli-Watt range [2]. The hardware cost can be further reduced by deploying single common RF power amplifier (PA) for a set of antenna elements and a separate RF phase shifter for each antenna, which is known as constant envelop precoding (CEP).

However, the price to pay for massive MIMO is increased complexity of the hardware (number of RF chains) and estimating channel coefficients for a large number of antennas. With the introduction of CEP, the hardware complexity can be greatly simplified [15] which we will elaborate in Section 8.3.2.

The lucrative features of massive MIMO described earlier have recently sparked an outbreak of research activities aimed at understanding the signal processing and information-theoretic upshots of massive MIMO system designs. In [2, 8], various aspects of massive MIMO systems are reviewed, including information-theoretic foundation, antenna and propagation aspects, channel acquisition, and transceiver design. More recently in [3], a follow-up survey briefly summarizes more recent works.

In this chapter, we provide a comprehensive foundation on massive MIMO signal processing and detailed overview of state-of-the-art research on this topic. Starting with an overview of multiantenna systems in Section 8.2, we show the information-theoretic path toward the development of massive MIMO technology. Various transmit precoding techniques are then discussed in Section 8.3, and signal detection schemes are presented in Section 8.4. Power control mechanisms for massive MIMO systems are summarized in Section 8.5. Although for the ease of exposition, the pre-coding/decoding schemes are discussed based on single-cell transmission, channel estimation, and pilot contamination issues are discussed from multicellular perspective in Section 8.6 for better understanding of a more realistic scenario. Finally, some challenging potential future research directions related to massive MIMO in future wireless communications are identified in Section 8.7.

---

## 8.2 OVERVIEW OF MULTIANTENNA SYSTEMS: PATH TO MASSIVE MIMO

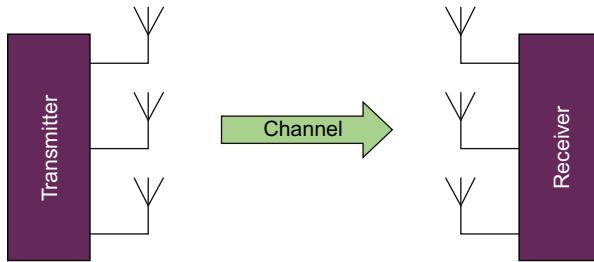
We start our discussion with the classical point-to-point MIMO link. However, the limitations of systems in which the working antennas are compactly clustered at the transmitter as well as the receiver lead naturally into the topic of multiuser MIMO (MU-MIMO). In fact, MU-MIMO is the ideal scenario where very large MIMO is envisioned to show its greatest efficacy. This section is based on the concepts developed in [2, 3, 16] to show the information-theoretic development toward massive MIMO. It has been shown that the Shannon theory simplifies greatly for large numbers of antennas at the BS and it suggests capacity-approaching solutions [2].

### 8.2.1 POINT-TO-POINT MIMO

In a point-to-point MIMO system, a transmitter equipped with an array of  $N_t$  antennas and a receiver equipped with an array of  $N_r$  antennas, as shown in Fig. 8.2, are connected by a wireless channel such that the signal received at every receive antenna is subject to the combined action of all transmit antennas. Considering the simplest narrowband channel, we have the received signal vector  $\mathbf{y} \in \mathbb{C}^{N_r \times 1}$  expressed as [17, 18]

$$\mathbf{y} = \sqrt{p}\mathbf{H}\mathbf{x} + \mathbf{n}, \quad (8.1)$$

where  $\mathbf{x} \in \mathbb{C}^{N_t \times 1}$  is the transmit signal vector,  $\mathbf{n} \in \mathbb{C}^{N_r \times 1}$  is the vector of noise and interference, and  $\mathbf{H} \in \mathbb{C}^{N_r \times N_t}$  is the propagation matrix of complex-valued channel

**FIG. 8.2**

A point-to-point MIMO system.

coefficients between transmit and receive antennas. The scalar  $p$  is the SNR of the link, which is proportional to the ratio of the transmitted power to the noise-variance. In what follows, it is assumed that the transmit signal  $\mathbf{x}$  is normalized such that the expected total transmit power is unity (i.e.,  $E\{\|\mathbf{x}\|^2\} = 1$ ). We further assume that the components of the additive noise vector  $\mathbf{n}$  are independent and identically distributed (i.i.d.) zero-mean and unit-variance circularly symmetric complex-Gaussian random variables ( $\mathcal{CN}(0, 1)$ ).

Assuming that the transmit signals are i.i.d. complex Gaussian random variables and that the receiver has perfect knowledge of the channel matrix,  $\mathbf{H}$ , the instantaneous mutual information (MI) between the transmitter and the receiver of the point-to-point MIMO channel (Eq. 8.1), measured in bits-per-channel-use, is given by [2, 3]

$$C = \log_2 \det \left( \mathbf{I}_{N_r} + \frac{p}{N_t} \mathbf{H} \mathbf{H}^H \right) \text{ bits s}^{-1} \text{ Hz}^{-1}. \quad (8.2)$$

The achievable rate of the channel approaches capacity if the input power is optimized according to the water-filling principle. If we assume that the propagation coefficients in the channel matrix are normalized such that  $\text{Tr}(\mathbf{H} \mathbf{H}^H) \approx N_t N_r$  and  $\mathbf{H} \mathbf{H}^H$  equals a scaled identity matrix, then  $C$  is in fact the capacity.

In order to dissect the MI expression in Eq. (8.2), we express the achievable rate in terms of the singular values of the propagation matrix,

$$\mathbf{H} = \mathbf{U}_h \boldsymbol{\Sigma}_h \mathbf{V}_h^H, \quad (8.3)$$

where  $\mathbf{U}_h$  and  $\mathbf{V}_h$  are  $N_r \times N_r$  and  $N_t \times N_t$  unitary matrices, respectively, and  $\boldsymbol{\Sigma}_h$  is an  $N_r \times N_t$  diagonal matrix whose diagonal elements are the singular values,  $\{\sigma_1, \sigma_2, \dots, \sigma_{\min(N_t, N_r)}\}$  of  $\mathbf{H}$ . Thus the achievable rate expression in Eq. (8.2) can be rewritten in terms of the singular values as

$$C = \sum_{l=1}^{\min(N_t, N_r)} \log_2 \left( 1 + \frac{p \sigma_l^2}{N_t} \right). \quad (8.4)$$

Clearly, the MI in Eq. (8.4) is equivalent to the combined achievable rate of parallel links for which the  $l$ th link has an SNR of  $\frac{p \sigma_l^2}{N_t}$ . Since the actual achievable rate

depends on the distribution of the singular values, let us now analyze the best case and the worst case scenario of the MIMO propagation channel. Subject to the constraint obtained from Eq. (8.3) that

$$\sum_{l=1}^{\min(N_t, N_r)} \sigma_l^2 = \text{Tr}(\mathbf{H}\mathbf{H}^H), \quad (8.5)$$

the worst case that may occur is only one nonzero singular value resulting in the lowest rate, and the best case is when all of the  $\min(N_t, N_r)$  singular values are equal, achieving the highest rate. These two cases actually bound the achievable rate (Eq. 8.4) as follows [2]:

$$\log_2(1 + pN_r) \leq C \leq \min(N_t, N_r) \log_2\left(1 + \frac{p \max(N_t, N_r)}{N_t}\right), \quad (8.6)$$

where the normalized channel matrix assumption  $\text{Tr}(\mathbf{H}\mathbf{H}^H) \approx N_t N_r$  has been utilized. However, in the low SNR regime where the system is power limited, having one large singular value than many smaller ones may produce better results.

There are many practical phenomena that may result in the worst case scenario including the compact arrays under line-of-sight (LOS) propagation conditions, or under extreme keyhole propagation conditions. The best case is approached when all of the entries in the channel matrix are distributed as i.i.d. random variables.

### 8.2.2 TOWARD MASSIVE MIMO

In this section, we focus on the emergence of massive MIMO based on the random matrix theory analysis [19]. Let us now consider two limiting cases, where either the number of transmit or the number of receive antennas increases without bounds.

- (1)  $N_t \gg N_r$  and  $N_t \rightarrow \infty$ : Let the number of transmit antennas grow large while keeping the number of receive antennas constant (i.e.,  $N_t \gg N_r, N_t \rightarrow \infty$ ). We further assume that the row-vectors of the propagation matrix  $\mathbf{H}$  are asymptotically orthogonal. Hence we have from [20]

$$\left(\frac{\mathbf{H}\mathbf{H}^H}{N_t}\right)_{N_t \gg N_r} \approx \mathbf{I}_{N_r}. \quad (8.7)$$

In this case, the achievable rate in Eq. (8.2) can be approximated as

$$C_{N_t \gg N_r} \approx N_r \log_2(1 + p) \text{ bits s}^{-1} \text{ Hz}^{-1}, \quad (8.8)$$

which achieves the rate upper bound in Eq. (8.6).

- (2)  $N_r \gg N_t$  and  $N_r \rightarrow \infty$ : Now, let the number of receive antennas grow large while keeping the number of transmit antennas constant (i.e.,  $N_r \gg N_t, N_r \rightarrow \infty$ ). We also assume that the column-vectors of the propagation matrix are asymptotically orthogonal, so

$$\left(\frac{\mathbf{H}^H \mathbf{H}}{N_r}\right)_{N_r \gg N_t} \approx \mathbf{I}_{N_t}. \quad (8.9)$$

Using the matrix identity  $\det(\mathbf{I} + \mathbf{A}\mathbf{A}^H) = \det(\mathbf{I} + \mathbf{A}^H\mathbf{A})$ , the achievable rate in Eq. (8.2) can be approximated as

$$C_{N_t \gg N_r} = \log_2 \det \left( \mathbf{I}_{N_t} + \frac{p}{N_t} \mathbf{H}^H \mathbf{H} \right) \approx N_t \log_2 (1+p) \text{ bits s}^{-1} \text{ Hz}^{-1}, \quad (8.10)$$

which achieves the rate upper bound in Eq. (8.6).

From the analysis earlier, we observe that an extreme number of transmit or receive antennas, combined with assumption that the row or column vectors of  $\mathbf{H}$  are asymptotically orthogonal, creates a highly desirable scenario. The key player is the excessive antennas that boost the effective SNR, which could even compensate for a low SNR and restore multiplexing gains that would otherwise be absent [2]. Furthermore, the orthogonality feature of the propagation vectors implies that i.i.d. complex Gaussian inputs are optimal so that the achievable rates given in Eqs. (8.8), (8.10) are in fact the actual channel capacities. This fundamental discovery sets the path toward massive MIMO.

### 8.2.3 MU-MIMO

The generous multiplexing gains offered by point-to-point MIMO, as discussed previously, require a favorable propagation environment. However, in the event of LOS propagation or when the terminal is at the cell edge, the achievable rate may not be satisfactory. Extra receive antennas can compensate for a low SNR, but that is not often a viable solution for mobile terminals. On the other hand, in MU-MIMO systems, terminals are generally separated by many wavelengths, and thus have the potential to obtain the promising multiplexing gain of massive point-to-point MIMO systems while eliminating problems due to unfavorable propagation environments [18, 21].

Let us now consider an MU-MIMO system in which a BS equipped with an array of  $N$  antennas serves  $K$  single-antenna users. Let  $h_{k,n}$  denote the channel coefficient from the  $k$ th user to the  $n$ th antenna of the BS, which is the product of a complex small-scale fading factor (i.e., which changes over intervals of a wavelength or less) and an amplitude factor that accounts for geometric attenuation and large-scale fading. Thus the channel coefficient can be expressed as

$$h_{k,n} = g_{k,n} \sqrt{d_k}, \quad (8.11)$$

where  $g_{k,n}$  and  $d_k$  represent complex small-scale fading and large-scale fading components, respectively. The large-scale fading accounts for path loss and shadowing effects. By assumption, the antenna array at the BS is sufficiently compact that the large-scale fading coefficients are the same for different antennas at the same BS, but are user-dependent. The small-scale fading coefficients are assumed to be different for different users or for different antennas at each BS. Thus the channel matrix from all  $K$  users to the BS can be expressed as

$$\mathbf{H} = \begin{pmatrix} h_{1,1} & \cdots & h_{K,1} \\ \vdots & \ddots & \vdots \\ h_{1,N} & \cdots & h_{K,N} \end{pmatrix} = \mathbf{G} \mathbf{D}^{1/2}, \quad (8.12)$$

where

$$\mathbf{G} = \begin{pmatrix} g_{1,1} & \cdots & g_{K,1} \\ \vdots & \ddots & \vdots \\ g_{1,N} & \cdots & g_{K,N} \end{pmatrix}, \quad \mathbf{D} = \begin{pmatrix} d_1 & & \\ & \ddots & \\ & & d_K \end{pmatrix}.$$

Thus, the  $k$ th column-vector of  $\mathbf{G}$  describes the small-scale fading between the  $k$ th terminal and the  $N$  antennas, while the  $k$ th diagonal element of  $\mathbf{D}^{1/2}$  is the corresponding large-scale fading coefficient. For convenience, the large-scale fading coefficients are often normalized such that the small-scale fading coefficients are typically of unit variances.

For typical MU-MIMO with large arrays, the number of BS antennas vastly exceeds the number of served terminals. As discussed previously, the column-vectors of the propagation matrix are asymptotically orthogonal under the most favorable propagation assumption, so that

$$\left( \frac{\mathbf{H}^H \mathbf{H}}{N} \right)_{N \gg K} = \mathbf{D}^{1/2} \left( \frac{\mathbf{G}^H \mathbf{G}}{N} \right)_{N \gg K} \mathbf{D}^{1/2} \approx \mathbf{D}. \quad (8.13)$$

Based on this theoretical foundation, we discuss both UL and DL signal models in the following sections.

### 8.2.3.1 UL (reverse link)

For UL signal transmission, the received signal vector  $\mathbf{y}_u \in \mathbb{C}^{N \times 1}$  at the BS can be expressed as

$$\mathbf{y}_u = \sqrt{p_u} \mathbf{H} \mathbf{x}_u + \mathbf{n}_u, \quad (8.14)$$

where  $\mathbf{x}_u \in \mathbb{C}^{K \times 1}$  is the signal vector transmitted from all the users,  $\mathbf{H} \in \mathbb{C}^{N \times K}$  is the UL channel matrix defined in Eq. (8.12),  $\mathbf{n}_u \in \mathbb{C}^{N \times 1}$  is a zero-mean noise vector with complex Gaussian distribution and identity covariance matrix, and  $p_u$  is the UL transmit power. The transmitted symbol from the  $k$ th user,  $x_{u,k}$ , is the  $k$ th element of  $\mathbf{x}_u = [x_{u,1}, \dots, x_{u,K}]^T$  with  $E[|x_{u,k}|^2] = 1$ .

Based on the assumption that the small-scale fading coefficients for different users are independent, the column channel vectors from different users are asymptotically orthogonal as the number of antennas at the BS,  $N$ , grows to infinity [5]. Then we have [20]

$$\mathbf{H}^H \mathbf{H} = \mathbf{D}^{1/2} \mathbf{G}^H \mathbf{G} \mathbf{D}^{1/2} \approx N \mathbf{D}^{1/2} \mathbf{I}_K \mathbf{D}^{1/2} = N \mathbf{D}. \quad (8.15)$$

Using Eq. (8.15) under perfect CSI assumption, the overall achievable rate of all users becomes

$$C_u = \log_2 \det(\mathbf{I}_K + p_u \mathbf{H}^H \mathbf{H}). \quad (8.16)$$

Note that the achievable sum rate of reverse link MU-MIMO is no less than that of the case as if the terminals could collaborate among themselves [17]. If collaboration were possible, it could definitely have made channel coding and decoding easier, but it would not alter the ultimate sum rate.

Based on the favorable propagation conditions obtained in Eq. (8.15), if there is a large number of antennas compared with the number of terminals, then the asymptotic sum rate is given by [2]

$$C_u \approx \log_2 \det(\mathbf{I}_K + Np_u \mathbf{D}) = \sum_{k=1}^K \log_2(1 + Np_u d_k) \text{ bits s}^{-1} \text{ Hz}^{-1}. \quad (8.17)$$

In the following, we show that with a very large antenna array, simple matched-filter (MF) processing at the BS can achieve the capacity in Eq. (8.17). With MF processing, the BS processes the received signal vector by multiplying the conjugate-transpose of the propagation channel. Thus the BS yields

$$\mathbf{H}^H \mathbf{y}_u = \mathbf{H}^H (\sqrt{p_u} \mathbf{H} \mathbf{x}_u + \mathbf{n}_u) \approx N \sqrt{p_u} \mathbf{D} \mathbf{x}_u + \mathbf{H}^H \mathbf{n}_u. \quad (8.18)$$

Note that the channel vectors are asymptotically orthogonal, that is,  $\mathbf{H}^H \mathbf{H} = N\mathbf{D}$  when the number of antennas at the BS grows to infinity. This has a nice intuition. Since  $\mathbf{D}$  is a diagonal matrix, the decoding of the transmission from the  $k$ th user requires only the  $k$ th component of Eq. (8.18). Thus the MF processing separates the signals from different users into different streams and the interuser interference disappears completely. This phenomenon can be interpreted as if the signal transmissions are originating from equivalent SISO channels. From Eq. (8.18), the SNR for the  $k$ th user is  $Np_u d_k$ . Accordingly, the rate achievable by using MF is the same as the limit in Eq. (8.17), which demonstrates that simple MF processing at the BS is optimal and capacity approaching when the number of antennas at the BS grows without limits.

### 8.2.3.2 DL (forward link)

Next, let us denote  $\mathbf{y}_d \in \mathbb{C}^{K \times 1}$  as the vector collecting the received signals at all the  $K$  users. For most prior works in massive MIMO, TDD mode is assumed as discussed in Section 8.6, where the DL channel is the transpose of the UL channel matrix. Then the received signal vector can be expressed as

$$\mathbf{y}_d = \sqrt{p_d} \mathbf{H}^T \mathbf{x}_d + \mathbf{n}_d, \quad (8.19)$$

where  $\mathbf{x}_d \in \mathbb{C}^{N \times 1}$  is the signal vector transmitted by the BS,  $\mathbf{n}_d \in \mathbb{C}^{K \times 1}$  is additive noise defined as before, and  $p_d$  is the transmit power of the DL. To normalize the transmit power, we assume  $E\{\|\mathbf{x}_d\|^2\} = 1$ . Thus the total transmit power is independent of the number of antennas.

In the literature of massive MIMO, it is generally assumed that the BS has CSI corresponding to all users based on UL pilot transmission exploiting channel reciprocity. Therefore, it is possible for the BS to perform power allocation to maximize the sum transmission rate. Let  $\mathbf{P}$  be a  $K \times K$  diagonal matrix with the power allocation vector  $[p_1, \dots, p_K]$  as its main diagonal and  $\sum_{k=1}^K p_k = 1$ . To obtain the sum-capacity with power allocation requires performing a constrained optimization [16],

$$\begin{aligned} C_d &= \max_{\mathbf{P}} \log_2 \det(\mathbf{I}_N + p_d \mathbf{H} \mathbf{P} \mathbf{H}^H) \\ \text{subject to } &\sum_{k=1}^K p_k = 1, \quad p_k \geq 0, \quad \forall k. \end{aligned} \quad (8.20)$$

Under favorable propagation conditions (8.15) and the assumption that  $N \gg K$ , the sum-capacity has a simple asymptotic form

$$\begin{aligned} C_{d,N \gg K} &= \max_{\mathbf{P}} \log_2 \det \left( \mathbf{I}_N + p_d \mathbf{P}^{1/2} \mathbf{H}^H \mathbf{H} \mathbf{P}^{1/2} \right) \\ &\approx \max_{\mathbf{P}} \log_2 \det (\mathbf{I}_K + p_d N \mathbf{P} \mathbf{D}) \text{ bits s}^{-1} \text{ Hz}^{-1} \\ &= \max_{\mathbf{P}} \sum_{k=1}^K \log_2 (1 + p_d N p_k d_k) \text{ bits s}^{-1} \text{ Hz}^{-1}. \end{aligned} \quad (8.21)$$

This result makes intuitive sense if the columns of the propagation matrix are nearly orthogonal, which occurs asymptotically as the number of antennas grows. Then the transmitter could use a simple MF linear precoder

$$\mathbf{x}_d = \frac{1}{\sqrt{N}} \mathbf{H}^* \mathbf{D}^{-1/2} \mathbf{P}^{1/2} \mathbf{s}_d, \quad (8.22)$$

where  $\mathbf{s}_d \in \mathbb{C}^{K \times 1}$  is the information symbol vector such that  $E\{|s_{d,k}|^2\} = 1$ . Then substitution of Eq. (8.22) into Eq. (8.19) yields the received signal vector at all  $K$  users as

$$\begin{aligned} \mathbf{y}_d &= \sqrt{p_d} \mathbf{H}^T \mathbf{H}^* \mathbf{D}^{-1/2} \mathbf{P}^{1/2} \mathbf{s}_d + \mathbf{n}_d \\ &\approx \sqrt{p_d N} \mathbf{D}^{1/2} \mathbf{P}^{1/2} \mathbf{s}_d + \mathbf{n}_d. \end{aligned} \quad (8.23)$$

Note that the approximation in Eq. (8.23) is based on the assumption that the number of antennas at the BS grows to infinity (i.e.,  $N \rightarrow \infty$ ). It can be easily observed that the overall achievable data rate in Eq. (8.23) can be maximized by optimal power allocation as in Eq. (8.21), which demonstrates that the MF precoder is capacity achieving in the DL as well.

### 8.3 MASSIVE MIMO PRECODING

In the previous section, we have demonstrated that based on the favorable propagation assumption of Eq. (8.15), the simple MF precoder/detector can achieve the capacity of a massive MU-MIMO system when the number of antennas at the BS,  $N$ , is much larger than the number of users,  $K$ , and grows to infinity (i.e.,  $N \gg K$  and  $N \rightarrow \infty$ ).

In order to design multiuser precoding in the forward link and detection in the reverse link, the BS needs CSI. In massive antenna array technology, it is widely considered that the users transmit pilot symbols simultaneously in the uplink and the BS constructs the channel estimates, which is subsequently used for precoding in the forward link based on channel reciprocity. In this section, we focus on various precoding schemes that can be adopted at the BS. We first discuss basic precoding methods and then extend the discussion to some practical issues related to massive MIMO precoding.

In regular MIMO systems with limited number of antennas, both linear and nonlinear precoding techniques have been considered viable [18, 22]. Compared

with linear precoding methods, nonlinear methods, such as dirty paper coding (DPC) [22, 23], vector perturbation (VP) [24], and lattice-aided methods [25], generally have superior performance although the implementation complexity significantly increases. However, with infinitely many antennas at the BS, nonlinear precoders are almost impractical in massive MIMO systems. By contrast, linear precoders, such as MF and ZF, have been proven to be near-optimal or as effective in most cases [2, 5]. Thus, it is more desirable to use linear precoding techniques in massive MIMO systems, which have significantly lower implementation complexity in practice. Therefore, we restrict our discussion on linear precoding techniques only in this section.

### 8.3.1 BASIC PRECODING SCHEMES

In what follows, we derive the precoders for a number of popular precoding techniques in the large antenna domain. The corresponding SNR/signal-to-interference-plus-noise ratio (SINR) expressions, with  $N, K \rightarrow \infty$ , but with a finite ratio  $\gamma = N/K$ , are tabulated in [Table 8.1](#) [2].

Let us first discuss the performance of a benchmark reference which is an interference-free (IF) system and may not even be practical. Obviously, the IF scheme should achieve the best performance that can be imagined since all the channel energy to terminal  $k$  is delivered to terminal  $k$  without causing any interuser interference. In this case, terminal  $k$  receives the sample

$$y_{d,k}^{\text{IF}} = \sqrt{\sum_{l=1}^N |h_{l,k}|^2 s_{d,k} + n_{d,k}}, \quad (8.24)$$

where  $s_{d,k}$  is the data symbol transmitted for user  $k$ . Since  $\sum_{l=1}^N |h_{l,k}|^2 / N \rightarrow 1, N \rightarrow \infty$ , and  $E\{|s_{d,k}|^2\} = p_d/K$ , the SNR per receiving unit for IF systems converges to  $p_d\gamma$  as  $N \rightarrow \infty$ .

Next, let us move on to practical precoding methods that are being widely considered in massive MIMO systems. The simplest approach is to invert the channel by means of the pseudo-inverse which is referred to as ZF precoding [17]. Variants of ZF include block diagonalization [26] and generalized ZF [27, 28], which are not covered in this article. Intuitively, when  $N$  grows to infinity,  $\mathbf{H}$  tends to have nearly orthogonal columns as the users are not correlated due to their geographical

**Table 8.1** SNR or SINR Expressions for Standard Precoding Techniques

Precoding Technique	SNR/SINR
Benchmark: Interference-free (IF)	$p_d\gamma$
Zero forcing (ZF)	$p_d(\gamma - 1)$
Matched filter (MF)	$\frac{p_d\gamma}{p_d + 1}$
Vector perturbation (VP)	$\approx \frac{p_d\gamma\pi}{6} \left(1 - \frac{1}{\gamma}\right)^{1-\gamma}, \gamma \leq 1.79$

locations. This in turn indicates that the performance of ZF precoding will be close to that of the benchmark IF system. However, a disadvantage of ZF is that processing cannot be done in a distributed manner at each antenna separately. Hence centralized processing is necessary [2].

Formally, with ZF precoding, the transmit signal from the BS can be expressed as

$$\mathbf{x}_d^{\text{ZF}} = \frac{1}{\sqrt{\alpha}} (\mathbf{H}^T)^\dagger \mathbf{s}_d = \frac{1}{\sqrt{\alpha}} \mathbf{H}^* (\mathbf{H}^T \mathbf{H}^*)^{-1} \mathbf{s}_d, \quad (8.25)$$

where the superscript  $\dagger$  indicates matrix pseudo-inverse, that is,  $(\mathbf{H}^T)^\dagger = \mathbf{H}^* (\mathbf{H}^T \mathbf{H}^*)^{-1}$ , and  $\alpha$  is the power normalization factor. The impact of normalization techniques has been discussed in [29]. A suitable choice of  $\alpha$  is  $\alpha = \text{Tr}(\mathbf{H}^T \mathbf{H}^*)^{-1}/K$ . Thus the received sample with ZF precoding becomes

$$y_{d,k}^{\text{ZF}} = \frac{1}{\sqrt{\alpha}} s_{d,k} + n_{d,k}, \quad (8.26)$$

which results in the instantaneous received SNR per terminal equal to

$$\text{SNR} = \frac{p_d}{K\alpha} = \frac{p_d}{\text{Tr}(\mathbf{H}^T \mathbf{H}^*)^{-1}}. \quad (8.27)$$

When both  $N$  and  $K$  grow large, but with a fixed ratio  $\gamma = N/K$ ,  $\text{Tr}(\mathbf{H}^T \mathbf{H}^*)^{-1}$  converges to a fixed deterministic value [2]

$$\text{Tr}(\mathbf{H}^T \mathbf{H}^*)^{-1} \rightarrow \frac{1}{\gamma - 1}. \quad (8.28)$$

Substituting Eq. (8.28) into Eq. (8.26) yields the corresponding expression in Table 8.1. An interesting insight from this analysis is that ZF precoding achieves an SNR that tends to the optimal SNR for an IF system with  $N - K$  transmit antennas when the array size increases.

For regularized ZF (RZF), a diagonal loading factor is added prior to the inversion of the matrix  $\mathbf{H}^T \mathbf{H}^*$ , and the transmit signal at the BS is expressed as [3]

$$\mathbf{x}_d^{\text{RZF}} = \frac{1}{\sqrt{\alpha}} \mathbf{H}^* (\mathbf{H}^T \mathbf{H}^* + \delta \mathbf{I})^{-1} \mathbf{s}_d, \quad (8.29)$$

where  $\delta > 0$  is the regularization factor that can be optimized based on the requirements of any particular system. The performance of ZF and RZF precoding has been studied in [30] for single-cell massive MIMO systems.

An inherent problem with ZF precoding is that the pseudo-inverse operation  $(\mathbf{H}^T)^\dagger = \mathbf{H}^* (\mathbf{H}^T \mathbf{H}^*)^{-1}$  requires the inversion of a  $K \times K$  matrix, which is computationally expensive. However, as  $N$  grows,  $(\mathbf{H}^T \mathbf{H}^*)/N$  approaches an identity matrix, which has a trivial inverse. Consequently, the ZF precoder tends to  $\mathbf{H}^*$ , which is in fact becoming an MF. This gives the intuition that matrix inversion may not be necessary when the antenna array is scaled up, as the MF precoder approximates the ZF precoder closely. For practical values of  $\gamma < \infty$ , the matrix can be simplified greatly. Thus applying MF, the transmit signal from the BS is expressed as

$$\mathbf{x}_d^{\text{MF}} = \frac{1}{\sqrt{\alpha}} (\mathbf{H}^T)^H \mathbf{s}_d = \frac{1}{\sqrt{\alpha}} \mathbf{H}^* \mathbf{s}_d, \quad (8.30)$$

with the normalization factor  $\alpha = \text{Tr}(\mathbf{H}^T \mathbf{H}^*)/K$ . A few simple mathematical derivations lead to an asymptotic expression of the SINR, which is given in Table 8.1. From the MF precoding SINR expression, it is evident that the SINR can be made as high as desired by scaling up the BS antenna array. However, the MF precoder exhibits an error floor since  $p_d \rightarrow \infty$  results in  $\text{SINR} \rightarrow \gamma$ . The impact of different normalization approaches is discussed in [29], which shows that vector normalization is better for ZF while matrix normalization is better for MF. Interestingly, the RZF precoder becomes the ZF precoder as  $\delta \rightarrow 0$ , and becomes the MF precoder as  $\delta \rightarrow \infty$ .

In Fig. 8.3, we show ergodic sum-rate capacities for MF precoding and ZF precoding based on the expressions in Table 8.1. As benchmark performance, we also plot the sum-rate capacity of an IF system. In all cases,  $K = 15$  users are considered and we show results for  $N = 15, 100, 500$ . In all cases, it can be seen that ZF decisively outperforms MF. However, as  $N$  grows, the advantage of IF quickly diminishes. With  $N = 100$ , the gain of IF is only about 0.5 dB. With 500 BS antennas, ZF precoding performs almost as good as an IF system.

The basic precoding schemes discussed earlier are based on the assumption that  $N, K \rightarrow \infty$ , but  $\gamma$  is finite. While the asymptotic analysis gives some basic insights about massive MIMO, any engineering work or optimization would likely use the finite- $N$  expressions. Hence in [14], the assumption has been restricted to a more realistic setting where  $N$  is not extremely large compared with  $K$  to analyze the extent to which the conclusions hold. In particular, the number of antennas required per UE has been determined to achieve  $\eta\%$  of the ultimate performance limit with infinitely many antennas. The results suggest that in certain scenarios, RZF/MMSE can perform as good as MF with almost one order of magnitude fewer antennas. The capacity bounds of the UL massive MIMO systems with finite number of BS antennas have been derived in [12], and that of the DL in [30] using ZF.

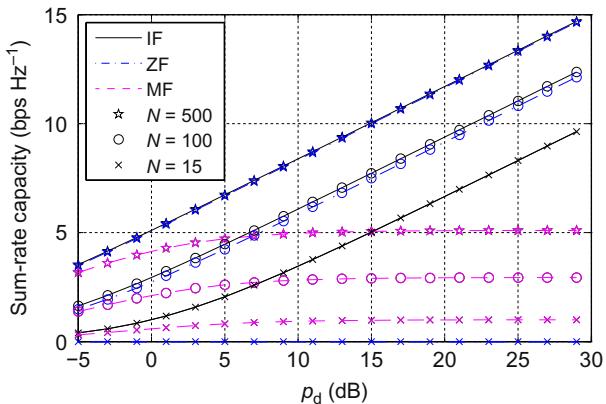


FIG. 8.3

Sum rates of single-cell MU-MIMO precoding techniques.

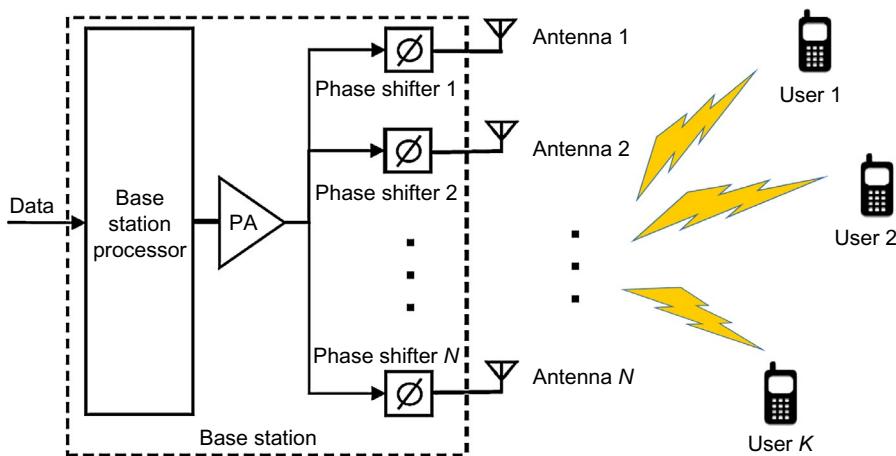
As discussed earlier, nonlinear precoding techniques, such as DPC, VP [24], and lattice-aided methods [25], produce practically interesting results when  $N$  is not much larger than  $K$ . However, the gap of ZF to an IF system scales as  $\gamma/(\gamma - 1)$ . Therefore, when  $N$  is, for example, two times of  $K$ , this gap is only 3 dB [2]. It is intuitive that nonlinear techniques will operate closer to the IF benchmark but cannot surpass it. Therefore, the gain of nonlinear methods in large antenna systems does not at all justify the complexity increase. For comparison, we include an approximate upper-bound SNR expression for VP, derived from the results of [31], in [Table 8.1](#). However, for larger values of  $\gamma$ , linear precoding performs well and the gain by using VP is not significant.

However, a practical challenge in the precoding design of massive multiuser MIMO systems is to facilitate hardware-friendly implementation. To this end, approximate message passing (AMP)-based precoding schemes have recently been proposed to minimize multiuser interference in massive MIMO systems [32]. Compared with the conventional precoding schemes, the AMP-based precoding is superior both in terms of computational complexity and average running time. This observation indicates that the AMP-based precoding is a suitable candidate for hardware implementation, which is very appealing for massive MIMO systems.

### 8.3.2 CONSTANT ENVELOPE PRECODING

In a precoder, weights of antenna elements are designed in a way to satisfy certain criteria such as interference power minimization. To implement the complex weights, each antenna element needs an RF PA and a phase shifter. Therefore, by increasing the number of antennas, the number of RF elements, and hence the implementation cost increases tremendously. One way to reduce this cost is to deploy a single common RF PA for all antenna elements and a separate RF phase shifter for each one. The system setup is shown in [Fig. 8.4](#). The scheme leads to antenna outputs with equal amplitudes but different phases. Precoding techniques exploiting this scenario are called CEP. Note that CEP was first considered in the context of massive MIMO in [15] by deploying one PA per antenna. The motivation for this was to facilitate the use of power-efficient PAs. In CEP, since antenna outputs have the same amplitudes and use the same PA, instead of linear amplifiers, more power efficient nonlinear amplifiers can be used [15]. Typically, a nonlinear PA can be about four to six times more power efficient than a linear one [33]. Therefore, with CEP, higher power efficiency and lower cost can outweigh the slight performance degradation, especially in the massive MIMO regime.

In [34], considering constant envelope transmission, authors derived the achievable rate of a multiinput single-output (MISO) Gaussian broadcast channel for the special case of a single user. They extended their work to the case of multiuser precoding in [15]. They were able to obtain the desired antenna phases by solving a nonlinear nonconvex continuous optimization problem with sum of squares of interuser interferences as the objective function to be minimized. It is shown that in the massive MIMO regime, the performance of their proposed CEP algorithm is comparable

**FIG. 8.4**

CEP in DL multiuser systems with a common power amplifier but separate phase shifters.

to that of precoding under an average only total transmit power constraint. To reduce the complexity, they also provide an alternating minimization method by taking all the phases but one to be a constant in each subiteration. That is, in each iteration, instead of solving an  $N$ -dimensional optimization problem, they solve  $N$  one-dimensional problems. The authors extended their work to frequency-selective channels in [35]. In [36], cross-entropy optimization method was used to solve the CEP problem, and it was shown that the resulting performance is less sensitive to the selection of the initial phase guesses.

Due to the practical limitations of phase shifters, in some cases, low-phase variations are desirable. To have small phase variations in each antenna element, in [13], the difference of the phase angles transmitted in consecutive channel uses is limited to a constant interval. The authors showed that in most cases, by achieving low-phase variations, the performance decreases by roughly 3 dB with every doubling in the number of BS antennas. In the existing literature, until now only analog phase shifters have been considered, which can support a continuous range of phase shifts. Nevertheless, digital phase shifters have certain additional benefits over analog phase shifters [37] as we will elaborate in [Section 8.7](#).

## 8.4 SIGNAL DETECTION

A fundamental aspect of MIMO technology is the computational complexity of the optimal detection, which increases exponentially with the problem size [17]. In principle the detection problem should scale up in the massive MIMO paradigm. However, thanks to a range of relevant results in random matrix theory, it is possible to keep massive MIMO detection relatively simpler [38]. As shown in [Section 8.2.3](#),

when the propagation matrix  $\mathbf{H}$  becomes larger, the channel becomes more and more deterministic, and its singular values become less sensitive to the actual distributions of the i.i.d. entries of  $\mathbf{H}$ . Eventually, the diagonal entries of  $\mathbf{H}^H\mathbf{H}$  become increasingly larger in magnitude than the off-diagonal entries. This behavior is often termed as *channel-hardening*, which is exploited for large-scale MIMO detection. As such, simple linear detectors, such as the MF, ZF, and minimum mean-square-error (MMSE) detectors, appear to be close to optimal if  $N \gg K$  under favorable propagation conditions [5, 39]. However, operating points with  $N \approx K$  are also important in practical systems with many users.

The performance of massive MIMO systems based on various linear receivers has been studied from various perspectives [14, 40–42]. A performance comparison between the MMSE and the MF receivers in realistic system settings is provided in [14]. The ergodic achievable rate as a function of the number of BS antennas has been provided. It shows that the MMSE receiver can achieve the same performance as the MF receiver with fewer antennas due to stronger multiuser interference. The scenario with a bounded ratio of the number of antennas to the number of users has been investigated in [40, 41] for the MMSE and ZF receivers, respectively. In [40], an expression for the asymptotic SINR of the MMSE receiver for a single-cell system with a bounded ratio of the number of antennas to the number of users is obtained. In [41], the exact data rate, symbol-error rate, and outage performance of the ZF receivers are derived. Besides centralized MIMO systems, the sum rate of the ZF receivers in distributed MIMO systems is also analyzed and lower and upper bounds on the sum rate are derived in [42]. The computational complexity of ZF and MMSE receivers is roughly in the order of  $\mathcal{O}(NK + NK^2 + K^3)$  [2].

Since pilot reuse affects the quality of massive MIMO channel estimation, robust receivers in a multiuser massive MIMO system against CSI mismatch caused by pilot contamination can further improve the performance of massive MIMO systems. In [43], a robust receiver design against pilot contaminated channel estimation is proposed by exploiting forward error correction code diversity as unique user signature to separate different pilot-interfering users during signal detection. Unlike the traditional detection approaches, the method in [43] distributes different channel codes among pilot-interfering users and utilizes the unique code signature to detect target user's signal, thereby achieving much better robustness against channel estimation error.

Apart from that, it has been shown recently that a widely linear ZF (WLZF) receiver with  $M$ -ASK modulation enjoys a spatial diversity gain, which linearly increases with the MIMO size [44]. In particular, the spatial diversity of WLZF under Rayleigh fading is  $N_r - N_t/2 + 1/2$  in contrast to the diversity of ZF, which is  $N_r - N_t + 1$ . The main advantage is that the WLZF receiver has low complexity with respect to the size of the massive MIMO system, even in very largely over-loaded scenarios. More recently, a near maximum likelihood detector for UL multiuser massive MIMO systems is proposed in [45] where each antenna is connected to a pair of 1-bit analog-to-digital converters (ADCs), one for each real and imaginary component of the baseband signal. Using the proposed solution in [45], the BS can perform simple symbol-by-symbol detection for the transmitted signals from multiple users.

Although practical transceiver implementations for wireless communication systems suffer from a number of hardware impairments that already occur at the transmit side, such as amplifier nonlinearities, quantization artifacts, and phase noise, the effect of such impairments has been routinely ignored in data detection. However, they often limit reliable communication in practice. A sophisticated data-detection algorithm has been presented in [46, 47], which takes into account a broad range of transmit-side impairments in large antenna systems. The key idea is to decouple the impaired MIMO system into  $N_t$  parallel and independent AWGN channels, which allows to perform impairment-aware maximum a posteriori data detection, independently for every user. The scheme has been extended in [48] to the case of mismatched data detection in massive MIMO systems, where the prior distribution of the transmit signal used in the data detector differs from the true prior.

## 8.5 POWER CONTROL

As mentioned in Section 8.2, appropriate power control in massive MIMO multicellular systems can further improve SE as well as EE. In this section, we discuss the recent developments in this area. Because most of the power in current networks is consumed at the BSs, the BS technology needs to be redesigned to reduce the power consumption as the wireless traffic grows. Existing results manifest that power control in massive MIMO systems can provide better SE for the users using less transmit power compared with low transmit power users in small-scale systems. In particular, Ref. [49] showed that most joint power allocation and BS-user association problems with power constraints are NP-hard. The recent work in [50] considered a relaxed problem formulation where each user can be associated with multiple BSs and showed that these problems can be solved by convex optimization.

However, the works in [49, 50] are all optimizing power with respect to the small-scale fading, which is very computationally demanding since the fading coefficients change rapidly (i.e., every few milliseconds). The power allocation can be optimized with respect to the slowly varying large-scale fading instead [8, 51–53], which makes advanced power control algorithms computationally feasible. A few recent works have considered power allocation for massive MIMO systems. For example, Ref. [51] formulated the DL EE optimization problem for the single cell massive MIMO systems that takes both the transmit and circuit powers into account. The paper [52] considered optimized user-specific pilot and data powers for given QoS constraints, while Ref. [53] optimized the max-min SE and sum SE. The SE-optimal scheme in [53] minimizes the total UL transmit power consumption when each user is served by an optimized subset of the BSs in a multicell massive MIMO system, using noncoherent joint transmission. A lower bound on the ergodic SE is derived, which is applicable for any channel distribution and precoding scheme. In addition, closed-form expressions are obtained for Rayleigh fading channels with either maximum ratio transmission (MRT) or ZF precoding.

In [54, 55], minimizing the total transmit power consumption was considered for multicell massive MIMO downlink systems when each user is served by the optimized subset of the BSs. A lower bound on the ergodic SE for Rayleigh fading channels and MRT was derived when the BSs cooperate using noncoherent joint transmission. The joint user association and downlink transmit power minimization problem was solved optimally under fixed SE constraints.

## 8.6 CHANNEL ESTIMATION AND PILOT CONTAMINATION

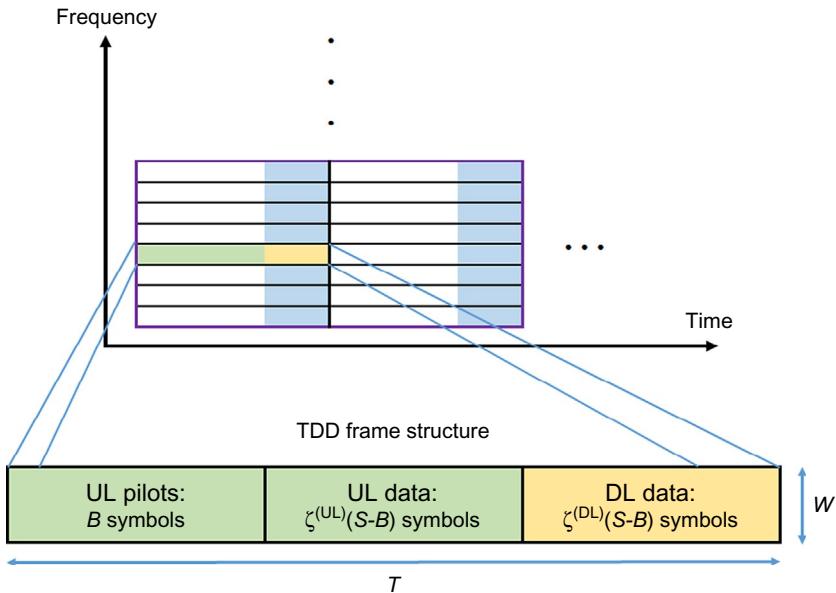
In this section, massive MIMO channel estimation and pilot contamination issues are discussed. We first discuss various channel estimation methods proposed for massive MIMO and explain why TDD mode is usually preferred. Then pilot contamination effect is elaborated.

### 8.6.1 CHANNEL ESTIMATION

In order to design multiuser precoding in the forward link and detection in the reverse link, the BS needs CSI. The CSI acquisition is normally performed by transmitting some pilot sequences. For regular MIMO systems, the time or frequency resources required for channel estimation are proportional to the number of the transmit antennas and are independent of the number of the receive antennas [56, 57].

In FDD, UL and DL use different frequency bands and hence the CSI corresponding to the UL and DL will be different. In the UL, CSI is estimated by allowing the users send different pilot sequences. Obviously, the time required for UL pilot transmission is independent of the number of antennas at the BS. However, estimating the CSI in the DL channel of FDD systems involves a two-stage procedure. To this end, the BS first transmits pilot symbols to all the users, and then all the users estimate the CSI and feed back to the BS via dedicated feedback channels. Hence the time required to transmit the DL pilot symbols is proportional to the number of antennas at the BS. As the number of BS antennas,  $N$ , grows large in massive MIMO systems, the traditional FDD channel estimation approach for the DL becomes infeasible. Let us assume that the frequency response of the propagation channel is constant over  $\eta$  consecutive subcarriers. The number of time slots devoted to orthogonal pilot transmission must be at least as large as  $N/\eta$ . With a limited coherence time in practice, when  $N$  grows infinitely, the time spent on transmitting pilots may surpass the coherence time of the channel. Eventually, the estimated DL CSI may become outdated. For example, a  $1\text{ ms} \times 100\text{ kHz}$  channel coherence interval can support transmission of 100 complex symbols. When there are 100 antennas at the BS (which may even be a few hundreds in practice), the whole coherence interval will be used for DL training if orthogonal pilot waveforms are used for channels to each antenna, while there would be no symbol time left for data transmission.

Fortunately, the channel estimation strategy in TDD systems can solve the problem identified earlier. Based on the assumption of channel reciprocity, only CSI for

**FIG. 8.5**

Massive MIMO TDD protocol.

the UL needs to be estimated. In [11], a TDD protocol, as shown in Fig. 8.5, was proposed. The time-frequency resources are divided into frames consisting of  $T$  seconds and  $W$  hertz. This allows transmission of  $S = TW$  symbols per frame. According to this protocol,  $B \geq 1$  out of the  $S$  symbols in each frame are reserved for UL pilot signaling. There is no DL pilot signaling and no feedback of CSI, because the BSs can process both UL and DL signals using the UL channel measurements due to the channel reciprocity in TDD systems. The remaining  $S - B$  symbols are allocated for payload data and are split between UL and DL transmission. Here  $\zeta^{(\text{ul})}$  and  $\zeta^{(\text{dl})}$  denote the fixed fractions allocated for UL and DL, respectively. These fractions can be selected arbitrarily, subject to the constraint  $\zeta^{(\text{ul})} + \zeta^{(\text{dl})} = 1$ . BSs use these pilot sequences to estimate CSI to the users located in their cells. Then the BSs use the estimated CSI to detect the UL data and generate beamforming vectors for DL data transmission. However, due to the limited channel coherence time, the pilot sequences used by users in neighboring cells may no longer be orthogonal to those within the cell, leading to a pilot contamination problem [5], which will be discussed in the next section. Because of pilot contamination, the estimated channel vector in any cell is a linear combination of channel vectors of users in other cells that use the same pilot sequence.

Linear MMSE-based channel estimation techniques are generally preferred for estimating CSI due to their near-optimal performance with low complexity. The exact closed-form expression for the MSE of the classical least square-based channel

estimation algorithm is derived in [58] to measure the pilot contamination in massive MIMO-OFDM systems. Based on that MSE expression, the optimal pilot design criterion has also been proposed. An MMSE-based channel estimation algorithm has been proposed in [59] for frequency selective channels that can achieve near optimal channel estimates at low complexity by exploiting the strong spatial correlation among antenna array elements. Stochastic geometry-based approaches have been adopted to quantify the pilot contamination, which is then utilized to analyze the effect of pilot contamination on the MSE of estimation.

Besides MMSE estimation, compressive sensing-based channel estimation approaches have also been proposed [60, 61]. Compressed sensing provides a framework for efficient CSI estimation utilizing prior knowledge of channel sparsity structures. It was shown in [62] for massive MIMO-OFDM systems that the common support of channel impulse responses is another source of utilizable sparsity structures. The impact of this support information on the required training overhead has been quantified in [63]. Other channel estimation techniques include tensor-based or parallel factor analysis-based methods [56, 57], which can perform even better than MMSE, but at the expense of increased complexity. In order to improve the SE in large-scale systems, a time-frequency-based approach is developed in [64]. The proposed estimation technique offers the benefits of both time- and frequency-domain estimation while eliminating their individual drawbacks. To exempt the fundamental assumption that the training duration should be relatively long to obtain acceptable CSI, Ref. [65] adopted a joint channel-and-data estimation method based on Bayes-optimal inference. This method yields minimal mean square errors with respect to the channels and payload data.

More recently, an ML channel estimation technique with one-bit ADCs that shares the same structure with the ML detector has been developed in [45]. The detectors and channel estimator in [45] provide a complete low-power channel estimation solution for the UL of a massive MIMO system. A joint channel estimation and decoding scheme has also been devised in [66] with the aid of the central limit argument and Taylor-series approximation for 3D massive MIMO-OFDM systems.

### 8.6.2 PILOT CONTAMINATION

Since pilot contamination is a dominating phenomenon that limits the performance of multicell multiuser massive MIMO systems, our discussion on this issue is based on multicell systems. Nonorthogonal pilot sequences used for channel estimation in multicell TDD networks is considered as a major source of pilot contamination in the literature due to limited coherence time [10]. In addition, the nonideal hardware components used in RF chains, which are often susceptible to impairments such as phase noise, amplifier nonlinearity, and quantization errors, may also lead to pilot contamination by affecting the accuracy of channel estimation [67].

As we have mentioned, TDD-based transmission is preferred for massive MIMO systems over its FDD counterpart due to the high signaling overhead and complexity associated with CSI acquisition under FDD. Hence we consider TDD case studies for

demonstrating pilot contamination effects. In TDD-based massive MIMO systems, pilot sequences are transmitted from users in the UL to acquire CSI. Let  $\boldsymbol{\pi}_{k,l} = [\pi_{k,l}^{[1]}, \dots, \pi_{k,l}^{[\tau]}]^T$  denote the pilot sequence transmitted by user  $k$  in cell  $l$ , where  $\tau$  is the length of the pilot sequence. The mathematical illustrations in this section follow the concepts demonstrated in [2, 3]. If orthogonal pilot sequences are transmitted by the users within the same cell as well as in the neighboring cells, then no contamination occurs, that is,

$$\boldsymbol{\pi}_{k,l}^H \boldsymbol{\pi}_{j,l'} = \delta[k-j] \delta[l-l'], \quad (8.31)$$

where  $\delta[\cdot]$  is defined as

$$\delta[n] = \begin{cases} 1 & n=0, \\ 0 & \text{otherwise.} \end{cases} \quad (8.32)$$

Under this pilot scheme, the estimated channel vectors are uncontaminated in the sense that they remain uncorrelated to the channel vectors of other users.

However, the limited number of orthogonal pilot sequences within a given time period and bandwidth eventually limits the number of users that can be concurrently served [3, 5]. In order to overcome that limit on users, nonorthogonal pilot sequences have to be used in neighboring cells such that for some different  $k, j, l$ , and  $l'$ , it may hold that  $\boldsymbol{\pi}_{k,l}^H \boldsymbol{\pi}_{j,l'} \neq 0$ . As a result, the estimated CSI of a user becomes correlated to that of the users with nonorthogonal pilot sequences.

With nonorthogonal pilot transmission, a key issue is to design efficient methods for distributing pilot sequences among the users in neighboring cells. Several schemes have been proposed for assigning nonorthogonal pilot sequences to users in different cells. One simple scheme that has been considered is to reuse the same set of orthogonal pilot sequences, say  $\{\boldsymbol{\pi}_1, \dots, \boldsymbol{\pi}_K\}$ , in all cells. This essentially means that the  $k$ th user in all cells will be assigned the pilot sequence  $\boldsymbol{\pi}_k$ . Duplicating pilot sequences in neighboring cells will eventually result in pilot contamination. During DL transmission, this contaminated CSI is used for precoder design. As a result, the BS steers transmission beams not only to users within the cell, but also to users in the neighboring cells, and therefore generates a strong directional interference. The problem is that, unlike the intracell interference, the interference due to pilot contamination will not disappear as the number of BS antennas increases. Thus one of the fundamental source of benefits of massive MIMO technology needs to be compromised. A similar effect can be observed during the UL transmission as well.

Let us now consider a UL massive MIMO system with  $L$  cells, in which each cell has  $K$  single-antenna users and a BS with  $N$  antennas, where  $N \gg K$ . For simplicity, it is assumed that all  $L$  cells use the same set of  $K$  pilot sequences, represented by the  $\tau \times K$  orthogonal matrix  $\boldsymbol{\Pi} = [\boldsymbol{\pi}_1, \dots, \boldsymbol{\pi}_K]$  such that  $\boldsymbol{\Pi}^H \boldsymbol{\Pi} = \tau \mathbf{I}_K$  holds. Let us further assume a synchronous pilot transmission from different cells. Thus the received signal matrix at the  $i$ th BS in the UL,  $\mathbf{Y}_i^p \in \mathbb{C}^N$ , can be expressed as

$$\mathbf{Y}_i^p = \sqrt{P_p} \sum_{l=1}^L \mathbf{H}_{i,l} \boldsymbol{\Pi}^T + \mathbf{N}_i^p, \quad (8.33)$$

where  $\mathbf{H}_{i,l} \in \mathbb{C}^{N \times K}$  denotes the channel matrix from all  $K$  users in the  $l$ th cell to the  $i$ th BS. The  $k$ th column of  $\mathbf{H}_{i,l}$ , denoted by  $\mathbf{h}_{i,k,l}$ , is the channel vector from the  $k$ th user in the  $l$ th cell to the  $i$ th BS. In addition,  $\mathbf{N}_i^p \in \mathbb{C}^{N \times \tau}$  is the noise matrix at the  $i$ th BS during the pilot transmission phase, whose entries are i.i.d. circular complex Gaussian random variables with zero-mean and unit variance, and  $p_p$  is the pilot transmit power.

In order to estimate the channel  $\mathbf{H}_{i,i}$ , the  $i$ th BS projects its received signal  $\mathbf{Y}_i^p$  on  $\Pi^*$ . Thus from Eq. (8.33), the estimated channel matrix is obtained as

$$\hat{\mathbf{H}}_{i,i} = \frac{1}{\sqrt{p_p \tau}} \mathbf{Y}_i^p \Pi^* = \mathbf{H}_{i,i} + \sum_{l \neq i} \mathbf{H}_{i,l} + \frac{1}{\sqrt{p_p \tau}} \mathbf{N}_i^p \Pi^*, \quad (8.34)$$

where  $\Pi^H \Pi = \tau \mathbf{I}_K$  has been used to obtain the second equality. The  $k$ th column,  $\hat{\mathbf{h}}_{i,k,i}$ , of  $\hat{\mathbf{H}}_{i,i}$  represents the estimate of the channel vector,  $\mathbf{h}_{i,k,i}$ . From Eq. (8.34), it can be easily observed that the estimate,  $\hat{\mathbf{h}}_{i,k,i}$ , is a linear combination of the channel vectors,  $\mathbf{h}_{i,k,i}$ , of the users in different cells with the same pilot sequence. In massive MIMO literature, this phenomenon is termed as pilot contamination effect [2, 3, 5].

After channel estimation, let us focus on the UL data transmission phase. Since we already know that the estimated CSI is contaminated, we can expect a resultant effect in signal detection. The received signal at the  $i$ th BS is given by

$$\mathbf{y}_i^u = \sqrt{p_u} \sum_{l=1}^L \sum_{k=1}^K \mathbf{h}_{i,k,l} x_{k,l}^u + \mathbf{n}_i^u, \quad (8.35)$$

where  $x_{k,l}^u$  is the symbol transmitted by the  $k$ th user in the  $l$ th cell and  $\mathbf{n}_i^u$  is the additive noise vector during UL transmission. Considering an MF decoder, the BS processes the signal vector by multiplying with the conjugate transpose of the estimated channel. Thus, the detected symbol from the  $k$ th user of the  $i$ th cell,  $\hat{x}_{k,i}^u$ , is found as

$$\hat{x}_{k,i}^u = (\hat{\mathbf{h}}_{i,k,i})^H \mathbf{y}_i^u = \left( \sum_{l=1}^L \mathbf{h}_{i,k,l} + \hat{\mathbf{v}}_i \right)^H \left( \sqrt{p_u} \sum_{l=1}^L \sum_{k=1}^K \mathbf{h}_{i,k,l} x_{k,l}^u + \mathbf{n}_i^u \right), \quad (8.36)$$

where  $\hat{\mathbf{v}}_i$  is the  $i$ th column of  $\frac{1}{\sqrt{p_p \tau}} \mathbf{N}_i^p \Pi^*$  in Eq. (8.34).

It can be observed from Eqs. (8.15), (8.36) that as the number of BS antennas grows large (i.e.,  $N \rightarrow \infty$ ), the received SINR of the  $k$ th user in the  $i$ th cell tends to the following limit [3, 5]

$$\text{SINR}_{k,i}^u = \frac{d_{i,k,i}^2}{\sum_{l \neq i} d_{i,k,l}^2}, \quad (8.37)$$

where  $d_{i,k,l}$  is the corresponding large-scale channel fading coefficient.

From Eq. (8.37), it is clear that the SINR depends only on the large-scale fading factors of the channels while the small-scale fading coefficients and noise just disappear. In addition, from Eq. (8.34), if nonorthogonal pilot sequences are used in different cells, the BS cannot distinguish among the channel vectors from its intracell

users to channels from users in other cells. The worst thing is that the interference due to pilot contamination will not disappear even though  $N$  is large enough compared to  $K$ . Similar results can be derived for the ZF and the MMSE detectors as well.

Since in TDD systems, DL precoding is designed based on the UL channel estimates, the pilot contamination phenomenon affects the DL transmission as well. For the DL, the power varies from one coherent interval to another if Eq. (8.34) is used directly as beamforming vectors. Thus, a normalized version for the beamforming vectors is commonly considered for better performance [3, 68]. The normalized MF beamforming vector from the  $i$ th BS to the  $k$ th user can be defined as

$$\mathbf{w}_{k,i}^d = \frac{\hat{\mathbf{h}}_{i,k,i}}{\|\hat{\mathbf{h}}_{i,k,i}\|} = \frac{\hat{\mathbf{h}}_{i,k,i}}{\alpha_{k,i}\sqrt{N}}, \quad (8.38)$$

where the scalar  $\alpha_{k,i} = \|\hat{\mathbf{h}}_{i,k,i}\| / \sqrt{N}$  is the normalization factor. Accordingly, the  $i$ th BS transmits an  $N$ -dimensional vector

$$\mathbf{x}_i^d = \sum_{k=1}^K \mathbf{w}_{k,i}^d s_{k,i}^d, \quad (8.39)$$

where  $s_{k,i}^d$  is the source symbol transmitted for the  $k$ th user. Hence the received signal at the  $k$ th user of the  $i$ th cell is expressed as

$$y_{k,i}^d = \sqrt{p_d} \sum_{l=1}^L \mathbf{h}_{l,k,i}^T \sum_{k'=1}^K (\mathbf{w}_{k',l}^d)^* s_{k',l}^d + n_{k,i}^d, \quad (8.40)$$

where  $n_{k,i}^d$  is the additive noise. Following similar derivation as for the UL, the DL SINR of the  $k$ th user in the  $i$ th cell, as  $N$  increases infinitely, tends to [3, 68]

$$\text{SINR}_{k,i}^d = \frac{\frac{d_{i,k,i}^2}{\alpha_{k,i}^2}}{\sum_{l \neq i} \frac{d_{l,k,i}^2}{\alpha_{k,l}^2}}, \quad (8.41)$$

where

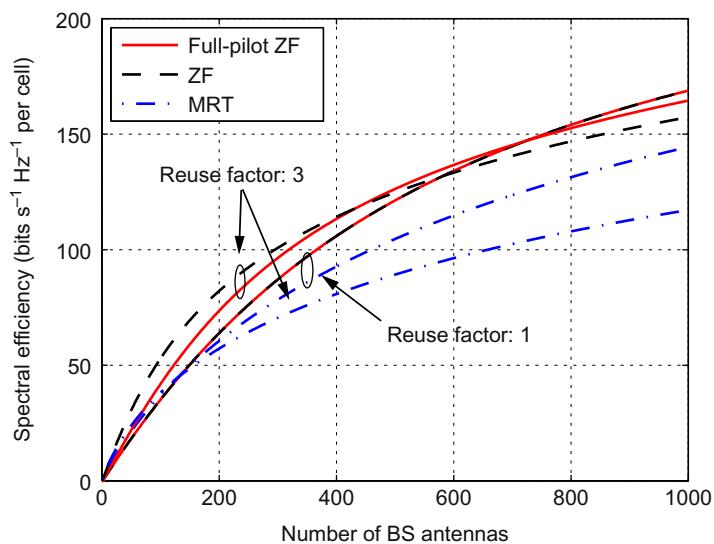
$$\alpha_{k,l}^2 = \sum_{j=1}^L d_{l,k,j} + \frac{1}{p_p \tau}. \quad (8.42)$$

Note that the statistical properties of the corresponding interference terms in the UL and the DL are different. In the UL, interference is generated by the reuse of the same pilot sequences among the users in different cells. In the DL, interference comes from the neighboring BSs that transmit using contaminated CSI. Even though the statistical properties are different for the UL and DL, the dissimilarities have little impact on the performance [3, 5, 68]. It is also important to note that payload data interference on the pilots has basically the same effect as pilot interference [12, Remark 5].

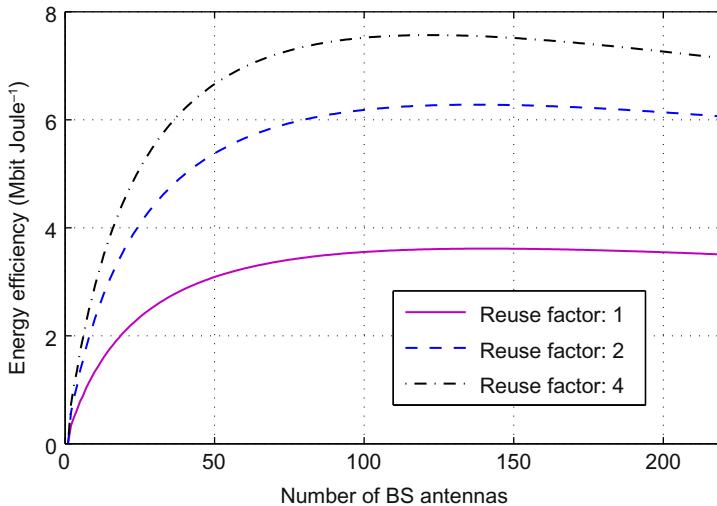
It is obvious from Eqs. (8.37), (8.41) that a massive MIMO system with strong pilot contamination has a lower performance limit than the one with weak pilot contamination. The review in [10] has analyzed the effect of pilot contamination on the

performance of massive MIMO systems, particularly on achievable rates. In [69], the convergence of SINR as the number of BS antennas goes to infinity with and without pilot contamination has been analyzed. The effect of SNR and the fading coefficients of the user channels on this rate of convergence have also been determined. The sum-rate lower bound in a pilot contaminated massive MIMO system is derived in [70], whereas the impact of practical channel models on the extent of pilot contamination is analyzed in [71]. An important aspect of the model in [71] is that it takes channel correlation of different users into account and is also applicable to scenarios without rich scattering. In addition, it has been shown that practical schemes like RZF/MMSE precoder/detector can perform as good as the MF precoder/detector with almost one order of magnitude fewer antennas even though the number of BS antennas is not extremely large compared to the number of users in each cell [14].

The effect of pilot contamination problem has also been demonstrated using different pilot reuse factors in multicell systems in the literature. It has been shown in [72] using frequency reuse factor of 1 that the sum rate in a massive MIMO system is much lower compared to higher reuse factors due to aggressive interference. Efficient system-level analyses have been performed in [11, 73] in order to derive SE and EE expressions, respectively, with arbitrary pilot reuse and random user locations. It has been demonstrated that up to half the coherence transmission block should be dedicated to pilots and the optimal  $N/K$  is less than 10 in many practical scenarios [11]. The maximal SE for different number of antennas is shown in Fig. 8.6, while Fig. 8.7 shows the maximum EE as a

**FIG. 8.6**

Effect of pilot reuse factors on massive MIMO SE.

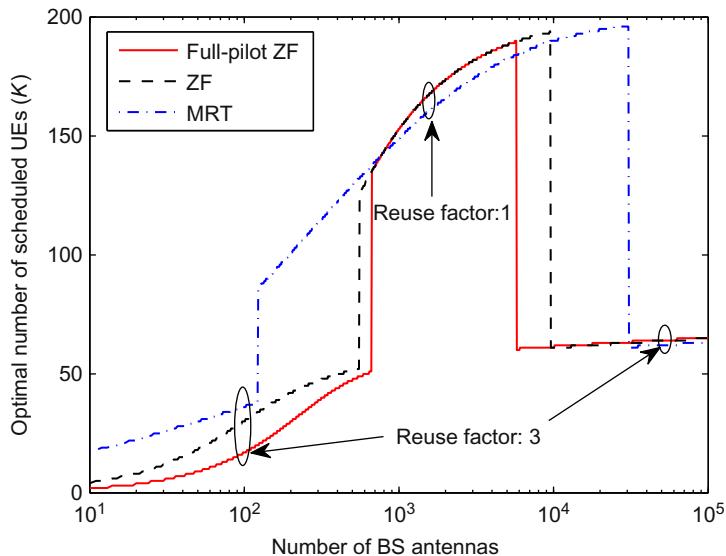
**FIG. 8.7**

Effect of pilot reuse factors on massive MIMO EE.

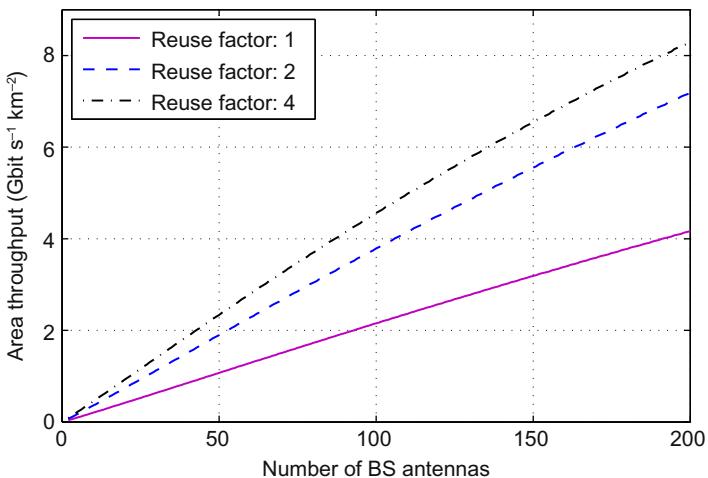
function of the number of BS antennas. The per-cell SEs shown in Fig. 8.6 provide the highest SEs for  $N \leq 1000$ , averaging over uniform UE locations in all cells. It can be observed that ZF brings much higher SEs than MRT, although the optimized SEs are rather similar for MRT, ZF, and full-pilot ZF in the practical range of  $10 \leq N \leq 200$  antennas. The largest differences appear when the number of antennas is very large. On the other hand, it is evident from Fig. 8.7 that pilot contamination affects massive MIMO systems by reducing the EE due to the intercell interference.

It has also been demonstrated in [11] that changes in the pilot reuse factor have major impact on the optimal number of UEs and their achievable performance. The optimal number of UEs that can be scheduled using different pilot reuse factors is plotted in Fig. 8.8, averaging over all cells. The average number of UEs and the per user throughput eventually defines the area throughput. An illustration of using different pilot reuse factor of 1, 2, 4 can be seen in Fig. 8.9 using Eq. (55) from [73] for area throughput. In Fig. 8.9, the use of pilot reuse factor of 1 shows a significant degradation in area throughput compared to reuse factors of 2 and 4 as a result of interference from all adjacent cells.

From the earlier results, an interesting observation is that the largest pilot reuse factor gives the highest EE (Fig. 8.7) and area throughput (Fig. 8.9). This demonstrates the necessity of actively mitigating pilot contamination in multicell massive MIMO systems.

**FIG. 8.8**

Optimal number of scheduled UEs in pilot contaminated massive MIMO systems.

**FIG. 8.9**

Effect of pilot contamination on massive MIMO throughput using different pilot reuse factors.

### 8.6.2.1 Mitigating pilot contamination effects

Attempts have been made to mitigate pilot contamination effects. In [72], an MMSE-based precoding is proposed that takes this particular form of the estimated channel vector into account and attempts to minimize (i) the sum of squares of errors for users

located in cell  $l$ ; and (ii) the sum of squares of interferences for users located in other cells. MMSE-based precoding provides significant improvement compared with traditional precoding methods like ZF, single-cell MMSE precoding [17]. In [74], the performance of massive MIMO ZF systems has been analyzed using the time-shifted pilot scheme, which was known to combat pilot contamination effectively using conjugate beamforming if there is a very large number of antennas. The authors in [75] proposed estimation of not only the channel parameters of the desired links in a target cell, but those of the interference links from adjacent cells. It has been shown that if the propagation properties of massive MIMO systems can be exploited, it is possible to obtain an accurate estimate of the channel parameters.

Prominent techniques considered in the literature have been identified in [3]. However, a comprehensive survey on established theories to mitigate pilot contamination, which include hardware impairment and nonreciprocal transceivers, has been provided in [10]. Here we briefly discuss some of them.

1. [(1)] Protocol-based methods: One way to reduce the pilot contamination effect is through frequency reuse or reducing the number of served users that use nonorthogonal pilot sequences [5]. However, using orthogonal pilots may not be guaranteed for the users scattered across different cells, because the convolution of the pilots with long channel impulse responses destroys their orthogonality. Moreover, the limited available bandwidth may not allow unique orthogonal pilots to be employed for each user. In this context, an effective pilot contamination elimination scheme has been proposed in [76] which can completely eliminate the pilot contamination effect. An optimal pilot design scheme has also been proposed in [77] based on time-shifted pilot transmission which is capable of completely eliminating pilot contamination under a much shorter coherence time.  
[(2)] Precoding methods: Specific precoding schemes can help mitigating pilot contamination effects. A distributed single-cell precoding method is proposed in [72] in which the precoding matrix at one BS is designed to minimize the sum of the squared error of its own users and interference to the users in all other cells. The authors demonstrated that the method can provide better performance compared to traditional single-cell ZF precoding.
2. The precoding method developed in [78] based on multicell cooperation can mitigate the pilot contamination effect. However, the required signaling overhead among the BSs increases with the number of antennas as well as BSs. Therefore, these methods are likely to find applications in limited MIMO scenarios. To obtain the benefit of cooperation while limiting the information exchange overhead, a pilot contamination precoding (PCP) method is proposed in [79]. It has been shown in [79] that PCP with properly optimized coefficients yields significantly large improvement over no-PCP scenario. For example, it has been demonstrated that at the outage probability 5%, optimal PCP provides 7500-fold gain in terms of achievable DL user rates.
3. Angle-of-arrival (AoA)-based methods: An interesting idea has also been demonstrated based on AoA. As shown in [80, 81], some users with identical or

nonorthogonal pilot sequences may suffer no interference at all with each other under realistic channel models. It has been proved in [80, 81] that users with mutually nonoverlapping AoA probability density functions hardly contaminate each other even if they use the same pilot sequence.

4. Blind methods: The blind methods work based on subspace partitioning. The methods are blind in the sense that they do not require pilot data to find the appropriate subspace. The methods follow the theory of large random matrices that predicts that the eigenvalue spectra of large sample covariance matrices can asymptotically decompose into disjoint bulks as the matrix size grows large. Random matrix and free probability theory are utilized to predict under which system parameters such a bulk decomposition takes place. The blind methods proposed for massive MIMO systems can also mitigate the effect of pilot contamination [82, 83].

## 8.7 FUTURE RESEARCH CHALLENGES

Based on our analysis so far, it is understandable that massive MIMO is a fast-developing plant in wireless communication research. However, there are still many issues that need to be addressed in order to harvest the maximum cropping from massive MIMO. In this section, we highlight only a few of the key challenges.

Pilot contamination is one of the inherent limitations of massive MIMO technology, which degrades the massive MIMO system performance significantly [2, 3, 84]. Significant amount of work focused on the impact of pilot contamination on system performance based on achievable rates, SE and EE. More work is expected on how pilot contamination influences the performance of the system based on quality of service. Furthermore, novel methods on mitigation of pilot contamination are expected for multicell systems that offer higher performance per users in terms of bit per second per hertz when a frequency reuse factor of 1 is employed. In addition, such methods are expected to take into consideration all sources of pilot contamination and not just limiting it to the nonorthogonal reuse training sequence. Evaluation of proposed methods based on complexity and cost is likely to pave way for new optimized methods in mitigating pilot contamination in TDD systems. Although the worst-case scenario have been extensively studied, the effect of pilot contamination in a dynamic or real-world system is likely to attract further studies by considering more deployment scenarios. Possible extension is to investigate the effects of pilot contamination in a multicell network by providing performance analysis based on Poisson point process and stochastic geometry. More research effort is required in mitigation of pilot contamination in the coexistence of massive MIMO in multilayer heterogeneous network (HetNet) where intra- and intertier interference is massive.

Another inherent complexity of massive MIMO is the massive number of RF chains at the BS potentially incurring high circuit power consumption, which would offset the benefits of “massive” in terms of EE [85]. Also, conventional MU-MIMO or massive MIMO may suffer from degradation in SNR at higher modulation orders.

Spatial modulation (SM) is a promising modulation scheme that could lead to increased SNR performance over conventional modulation schemes on the order of several decibels with the same SE in large-scale MU-MIMO systems [86–88]. In addition, the underlying technology can alleviate the requirement of large number of transmit RF chains in massive MIMO systems since in a given channel use, only one of the  $N$  transmit antennas will be activated, and the remaining  $N - 1$  transmit antennas will remain silent. In addition to the information bits conveyed through conventional modulation symbols on the active antenna, the index of the active transmit antenna also conveys information bits [86–88]. Hence, the number of bits conveyed in one channel use in SM is  $\lfloor \log_2 N \rfloor + \lfloor \log_2 C_\alpha \rfloor$  where  $C_\alpha$  is the cardinality of the modulation alphabet [87]. Exploiting SM, a compressive sensing-based detector has been designed in [61] that allows the reduction of the signal processing load at the BSs particularly pronounced for massive MIMO systems. Although there have been some progress and initial achievements in this potential research direction, key techniques and performance indicators are still unexplored. In particular, the optimal SM capacity bounds in large-scale antenna systems are yet unknown. Hence considerable attention is needed in this direction.

As discussed earlier, in the existing literature until now, only analog phase shifters have been considered for CEP, which can support a continuous range of phase shifts. However, in practice, due to advantages like increased immunity to noise on their voltage control lines, more uniform unit-to-unit performance and having flat phase over a wider bandwidth, mostly digital phase shifters are used [37]. Digital phase shifters provide a discrete set of phase states, which are controlled by a string of binary digits; the more the number of supported phase states, the higher the price. For instance, a 2-bit digital phase shifter supports only four phase states, namely, 0 degree, 90 degrees, 180 degrees, and 270 degrees [37]. Therefore, assuming practical digital phase shifters, the design problem for CEP turns into a nonlinear discrete optimization problem. In general, discrete optimization problems are NP-hard, so they are usually solved using some suboptimal algorithms which in worst-case scenarios (depending on the problem parameters) can have exponential time complexity [89]. Even in well-conditioned problems, when the dimension of the problem grows, as in the case of massive MIMO systems, their solution time is nontrivial and in most cases impractical. On the other hand, using CEP with discrete phases, interference powers at the users can be made arbitrarily small by taking the number of BS antennas to be sufficiently large. Hence, low-complexity algorithms must be developed to find the optimal phases to be used in order to harvest the full benefits of discrete-time (digital) phase shifters. Optimal tradeoff between the number of phases and complexity must also be addressed. The effects of phase and amplitude errors in practical phase shifters on the performance of CEP algorithms also need to be analyzed.

Another potential technique has been proposed for reducing the massive MIMO hardware costs based on electromagnetic (EM) lens-focusing antenna arrays [90, 91]. The EM lens has the capability of focusing the power of an incident wave to a small area of the antenna array, whereas the location of the focal area varies with

the AoA of the wave. Hence, in scenarios where the arriving signals from geographically separated users have different AoAs, the EM lens-enabled receiver can provide additional benefits. However, significant works still need to be done in order to define the performance bounds.

In addition, massive MIMO systems operating in the millimeter wave (mmWave) band can achieve orders of magnitude increase in SE and EE, which usually exploits the hybrid analog and digital precoding to overcome the serious signal attenuation induced by mmWave frequencies. In hybrid precoding schemes, the analog precoder is selected from the antenna array response vectors while the digital precoder is chosen from a predefined codebook [92–94]. Recently, there has been growing interest in hybrid beamforming design for large-scale mmWave communication. However, there are many challenging signal processing issues regarding the deployment of hybrid beamforming in practice [93]. Before implementing in practice, the performance gap compared to digital beamforming needs to be identified. Also, the EE-SE optimal bounds are still unknown. The design of efficient reference signals for better CSI estimation at the transmitter side is also an open problem.

Most of the existing works in the literature considered massive MIMO with collocated antenna arrays at the BS. However, massive number of antennas distributed over a large area can also form a massive MIMO array. Design and analysis of distributed massive MIMO systems are of practical interest since the associated processing task can be vastly shouldered by the distributed nodes. Some related works can be found in the massive MIMO literature. For example, it was shown in [95] by clustering the cooperating BSs and partitioning the users into groups that distributed massive MIMO schemes can achieve an SE comparable with that of collocated massive MIMO in [5] with a much smaller number of active antenna elements. However, in [5], the authors considered conjugate beamforming. In future work, the schemes in [5, 95] could be considered with ZF beamforming. It is also interesting to compare the EE between the system in [95] and the system in [5].

The acquisition of CSI is another challenging task in massive MIMO systems, in particular, when the number of BS antennas grow very large. Although significant research has been done in this direction, there are still many questions not appropriately answered [84]: Can we estimate the channels blindly? Can we utilize the payload data in estimating the channels in order to improve accuracy? Should each user estimate the effective channel in the DL instead? Is the gain obtained from such schemes significant enough to justify the associated costs?

---

## REFERENCES

- [1] E. Hossain, M. Hasan, 5G cellular: key enabling technologies and research challenges, *IEEE Instrum. Meas. Mag.* 18 (2015) 11–21.
- [2] F. Rusek, D. Persson, B.K. Lau, E.G. Larsson, T.L. Marzetta, O. Edfors, F. Tufvesson, Scaling up MIMO: opportunities and challenges with very large arrays, *IEEE Signal Process. Mag.* 30 (2013) 40–60.
- [3] L. Lu, G.Y. Li, A.L. Swindlehurst, A. Ashikhmin, R. Zhang, An overview of massive MIMO: benefits and challenges, *IEEE J. Sel. Top. Signal Process.* 8 (2014) 742–758.

- [4] T.L. Marzetta, Multi-cellular wireless with base stations employing unlimited numbers of antennas, in: Proc. UCSD Inf. Theory Applications Workshop, 2010.
- [5] T.L. Marzetta, Noncooperative cellular wireless with unlimited numbers of base station antennas, *IEEE Trans. Wirel. Commun.* 9 (2010) 3590–3600.
- [6] E. Björnson, E.G. Larsson, T.L. Marzetta, Massive MIMO: ten myths and one critical question, *IEEE Commun. Mag.* 54 (2016) 114–123.
- [7] S. Malkowsky, K. Nieman, Z. Miers, N. Kundargi, L. Liu, I. Wong, V. Öwall, O. Edfors, F. Tufvesson, A flexible 100-antenna testbed for massive MIMO, in: IEEE Globecom Workshops (GC Wkshps), Austin, TX, 2014, pp. 287–293.
- [8] E.G. Larsson, F. Tufvesson, O. Edfors, T.L. Marzetta, Massive MIMO for next generation wireless systems, *IEEE Commun. Mag.* 52 (2014) 186–195.
- [9] T.L. Marzetta, Massive MIMO: an introduction, *Bell Labs Tech. J.* 20 (2015) 11–22.
- [10] O. Elijah, C.Y. Leow, T.A. Rahman, S. Nunoo, S.Z. Iliya, A comprehensive survey of pilot contamination in massive MIMO-5G system, *IEEE Commun. Surv. Tutorials* 18 (2016) 905–923.
- [11] E. Björnson, E.G. Larsson, M. Debbah, Massive MIMO for maximal spectral efficiency: how many users and pilots should be allocated? *IEEE Trans. Wirel. Commun.* 15 (2016) 1293–1308.
- [12] H.Q. Ngo, E.G. Larsson, T.L. Marzetta, Energy and spectral efficiency of very large multiuser MIMO systems, *IEEE Trans. Commun.* 61 (2013) 1436–1449.
- [13] S. Mukherjee, S.K. Mohammed, Constant-envelope precoding with time-variation constraint on the transmitted phase angles, *IEEE Wirel. Commun. Lett.* 4 (2015) 221–224.
- [14] J. Hoydis, S. ten Brink, M. Debbah, Massive MIMO in the UL/DL of cellular networks: how many antennas do we need? *IEEE J. Sel. Areas Commun.* 31 (2013) 160–171.
- [15] S.K. Mohammed, E.G. Larsson, Per-antenna constant envelope precoding for large multi-user MIMO systems, *IEEE Trans. Commun.* 61 (2013) 1059–1071.
- [16] S. Vishwanath, N. Jindal, A. Goldsmith, Duality, achievable rates, and sum-rate capacity of Gaussian MIMO broadcast channels, *IEEE Trans. Inf. Theory* 49 (2003) 2658–2668.
- [17] D. Tse, P. Viswanath, *Fundamentals of Wireless Communication*, Cambridge University Press, Cambridge, 2005.
- [18] M.R.A. Khandaker, Y. Rong, Joint transceiver optimization for multiuser MIMO relay communication systems, *IEEE Trans. Signal Process.* 60 (2012) 5977–5986.
- [19] H. Cramér, *Random Variables and Probability Distributions*, Cambridge University Press, Cambridge, 1970.
- [20] M. Matthaiou, M.R. McKay, P.J. Smith, J.A. Nossek, On the condition number distribution of complex Wishart matrices, *IEEE Trans. Commun.* 58 (2010) 1705–1717.
- [21] M.R.A. Khandaker, Y. Rong, Joint source and relay optimization for multiuser MIMO relay communication systems, in: Proc. 4th Int. Conf. Signal Process. Commun. Systems (ICSPCS’2010), Gold Coast, Australia, December 13–15, 2010.
- [22] M.R.A. Khandaker, Y. Rong, Dirty paper coding based optimal MIMO relay communications, in: Proc. 16th Asia-Pacific Conf. Commun. (APCC’2010), Auckland, New Zealand, November 1–3, 2010, pp. 291–296.
- [23] M.H.M. Costa, Writing on dirty paper, *IEEE Trans. Inf. Theory* 29 (1983) 439–441.
- [24] B.M. Hochwald, C.B. Peel, A.L. Swindlehurst, A vector-perturbation technique for near-capacity multiantenna communication, part II: perturbation, *IEEE Trans. Commun.* 53 (2005) 537–544.
- [25] C. Windpassinger, R.F.H. Fischer, J.B. Huber, Lattice-reduction-aided broadcast precoding, *IEEE Trans. Commun.* 52 (2004) 2057–2060.

- [26] L.-U. Choi, R.D. Murch, A transmit preprocessing technique for multiuser MIMO systems using a decomposition approach, *IEEE Trans. Wirel. Commun.* 3 (2004) 20–24.
- [27] K.K. Wong, R.D. Murch, K. Ben Letaief, A joint-channel diagonalization for multiuser MIMO antenna systems, *IEEE Trans. Wirel. Commun.* 4 (2003) 773–786.
- [28] Z.G. Pan, K.K. Wong, T.S. Ng, Generalized multiuser orthogonal space division multiplexing, *IEEE Trans. Wirel. Commun.* 3 (2004) 1–5.
- [29] C. Lee, C.B. Chae, T. Kim, S. Choi, J. Lee, Network massive MIMO for cell-boundary users: from a precoding normalization perspective, in: Proc. IEEE Globecom Workshops, Anaheim, CA, 2012, pp. 233–237.
- [30] H. Yang, T.L. Marzetta, Performance of conjugate and zero-forcing beamforming in large-scale antenna systems, *IEEE J. Sel. Areas Commun.* 31 (2013) 172–179.
- [31] D.J. Ryan, I.B. Collings, I.V.L. Clarkson, R.W. Heath, Performance of vector perturbation multiuser MIMO systems with limited feedback, *IEEE Trans. Commun.* 57 (2009) 2633–2644.
- [32] J.C. Chen, C.J. Wang, K.-K. Wong, C.K. Wen, Low-complexity precoding design for massive multiuser MIMO systems using approximate message passing, *IEEE Trans. Veh. Technol.* 65 (2016) 5707–5714.
- [33] S.C. Cripps, *RF Power Amplifiers for Wireless Communications*, second ed., Artech House, Norwood, MA, 2006.
- [34] S.K. Mohammed, E.G. Larsson, Single-user beamforming in large-scale MISO systems with per-antenna constant-envelope constraints: the doughnut channel, *IEEE Trans. Wirel. Commun.* 11 (2012) 3992–4005.
- [35] S.K. Mohammed, E.G. Larsson, Constant-envelope multi-user precoding for frequency-selective massive MIMO systems, *IEEE Wirel. Commun. Lett.* 2 (2013) 547–550.
- [36] J.-C. Chen, C.-K. Wen, K.-K. Wong, Improved constant envelope multiuser precoding for massive MIMO systems, *IEEE Commun. Lett.* 18 (2014) 1311–1314.
- [37] Microwave101, Phase shifters, Available from: <http://www.microwaves101.com/encyclopedia/phaseshifters>.
- [38] A.M. Tulino, S. Verdú, Random matrix theory and wireless communications, *Found. Trends Commun. Inf. Theory* 1 (2004) 1–182.
- [39] S. Yang, L. Hanzo, Fifty years of MIMO detection: the road to large-scale MIMOs, *IEEE Commun. Surv. Tutorials* 17 (2015) 1941–1988.
- [40] Y.-C. Liang, G.M. Pan, Z.D. Bai, Asymptotic performance of MMSE receivers for large systems using random matrix theory, *IEEE Trans. Inf. Theory* 53 (2007) 4173–4190.
- [41] H.Q. Ngo, M. Matthaiou, T.Q. Duong, E.G. Larsson, Uplink performance analysis of multiuser MU-SIMO systems with ZF receivers, *IEEE Trans. Veh. Technol.* 62 (2013) 4471–4483.
- [42] M. Matthaiou, C. Zhong, M.R. McKay, T. Ratnarajah, Sum rate analysis of ZF receivers in distributed MIMO systems, *IEEE J. Sel. Areas Commun.* 31 (2013) 180–191.
- [43] K. Wang, Z. Ding, Robust receiver design based on FEC code diversity in pilot-contaminated multi-user massive MIMO systems, in: Proc. 2016 IEEE ICASSP, Shanghai, China, March 20–25, 2016, pp. 3426–3430.
- [44] J.C. De Luna Ducoing, N. Yi, Y. Ma, R. Tafazolli, Using real constellations in fully- and over-loaded large MU-MIMO systems with simple detection, *IEEE Wirel. Commun. Lett.* 5 (2016) 92–95.
- [45] J. Choi, J. Mo, R.W. Heath, Near maximum-likelihood detector and channel estimator for uplink multiuser massive MIMO systems with one-bit ADCs, *IEEE Trans. Commun.* 64 (2016) 2005–2018.

- [46] C. Jeon, R. Ghods, A. Maleki, C. Studer, Optimality of large MIMO detection via approximate message passing, in: Proc. IEEE ISIT, Hong Kong, June, 2015, pp. 1227–1231.
- [47] R. Ghods, C. Jeon, A. Maleki, C. Studer, Optimal large-MIMO data detection with transmit impairments, in: Proc. 53rd Allerton Conf. Commun. Control Comput., Monticello, IL, 2015, pp. 1211–1218.
- [48] C. Jeon, A. Maleki, C. Studer, On the performance of mismatched data detection in large MIMO systems, in: Proc. IEEE Int. Symp. Inf. Theory (ISIT), Barcelona, July 10–15, 2016, pp. 180–184.
- [49] R. Sun, M. Hong, Z.Q. Luo, Joint downlink base station association and power control for max-min fairness: computation and complexity, *IEEE J. Sel. Areas Commun.* 33 (2015) 1040–1054.
- [50] J. Li, E. Björnson, T. Svensson, T. Eriksson, M. Debbah, Joint precoding and load balancing optimization for energy-efficient heterogeneous networks, *IEEE Trans. Wirel. Commun.* 14 (2015) 5810–5822.
- [51] L. Zhao, H. Zhao, F. Hu, K. Zheng, Z. Zhang, Energy efficient power allocation algorithm for downlink massive MIMO with MRT precoding, in: Proc. IEEE VTC-Fall, 2013.
- [52] K. Guo, Y. Guo, G. Fodor, G. Ascheid, Uplink power control with MMSE receiver in multi-cell MU-massive-MIMO systems, in: Proc. IEEE ICC, 2014, pp. 5184–5190.
- [53] H.V. Cheng, E. Björnson, E.G. Larsson, Uplink pilot and data power control for single cell massive MIMO systems with MRC, in: Proc. Int. Symp. Wirel. Commun. Systems (ISWCS), Brussels, 2015, pp. 396–400.
- [54] T. Van Chien, E. Björnson, E.G. Larsson, Downlink power control for massive MIMO cellular systems with optimal user association, in: Proc. IEEE ICC, Kuala Lumpur, Malaysia, 2016.
- [55] T.V. Chien, E. Björnson, E. Larsson, Joint power allocation and user association optimization for massive MIMO systems, *IEEE Trans. Wirel. Commun.* 15 (2016) 6384–6399.
- [56] Y. Rong, M.R.A. Khandaker, Channel estimation of dual-hop MIMO relay system via parallel factor analysis, in: Proc. 17th Asia-Pacific Conf. Commun. (APCC’2011), Sabah, Malaysia, October 2–5, 2011.
- [57] Y. Rong, M.R.A. Khandaker, Y. Xiang, Channel estimation of dual-hop MIMO relay systems using parallel factor analysis, *IEEE Trans. Wirel. Commun.* 11 (2012) 2224–2233.
- [58] S. Ni, J. Zhao, R. Ran, Analysis of channel estimation in large-scale MIMO aided OFDM systems with pilot design, in: Proc. IEEE 83rd Veh. Technol. Conf. (VTC Spring), Nanjing, China, 2016.
- [59] A. Zaib, M. Masood, A. Ali, W. Xu, T.Y. Al Naffouri, Distributed channel estimation and pilot contamination analysis for massive MIMO-OFDM systems, *IEEE Trans. Commun.* 64 (11) (2016) 4607–4621.
- [60] S. Nguyen, A. Ghayeb, Compressive sensing-based channel estimation for massive multiuser MIMO systems, in: Proc. IEEE Wirel. Commun. Netw. Conf., Shanghai, China, April, 2013, pp. 2890–2895.
- [61] A. Garcia-Rodriguez, C. Masouros, Low-complexity compressive sensing detection for spatial modulation in large-scale multiple access channels, *IEEE Trans. Commun.* 63 (2015) 2565–2579.
- [62] C. Qi, Y. Huang, S. Jin, L. Wu, Sparse channel estimation based on compressed sensing for massive MIMO systems, in: Proc. IEEE Int. Conf. Commun. (ICC), 2015, pp. 4558–4563.

- [63] J.C. Shen, J. Zhang, E. Alsusa, K.B. Letaief, Compressed CSI acquisition in FDD massive MIMO: how much training is needed? *IEEE Trans. Wirel. Commun.* 15 (2016) 4145–4156.
- [64] L. Dai, Z. Wang, Z. Yang, Spectrally efficient time-frequency training OFDM for mobile large-scale MIMO systems, *IEEE J. Sel. Areas Commun.* 31 (2013) 251–263.
- [65] C.K. Wen, C.J. Wang, S. Jin, K.K. Wong, P. Ting, Bayes-optimal joint channel-and-data estimation for massive MIMO with low-precision ADCs, *IEEE Trans. Signal Process.* 64 (2015) 2541–2556.
- [66] S. Wu, L. Kuang, Z. Ni, D. Huang, Q. Guo, J. Lu, Message-passing receiver for joint channel estimation and decoding in 3D massive MIMO-OFDM systems, *IEEE Trans. Wirel. Commun.* 15 (12) (2016) 8122–8138.
- [67] E. Björnson, J. Hoydis, M. Kountouris, M. Debbah, Massive MIMO systems with non-ideal hardware: energy efficiency, estimation, and capacity limits, *IEEE Trans. Inf. Theory* 60 (2014) 7112–7139.
- [68] F. Fernandes, A. Ashikhmin, T.L. Marzetta, Inter-cell interference in noncooperative TDD large scale antenna systems, *IEEE J. Sel. Areas Commun.* 31 (2013) 192–201.
- [69] B. Gopalakrishnan, N. Jindal, An analysis of pilot contamination on multi-user MIMO cellular systems with many antennas, in: Proc. Signal Process. Adv. Wirel. Commun. (SPAWC), San Francisco, CA, USA, 2011, pp. 381–385.
- [70] D. Wang, C. Ji, X. Gao, S. Sun, X. You, Uplink sum-rate analysis of multi-cell multi-user massive MIMO system, in: Proc. IEEE Int. Conf. Commun. (ICC), Budapest, Hungary, June, 2013, pp. 5404–5408.
- [71] H.Q. Ngo, E.G. Larsson, T.L. Marzetta, The multicell multiuser MIMO uplink with very large antenna arrays and a finite-dimensional channel, *IEEE Trans. Commun.* 61 (2013) 2350–2361.
- [72] J. Jose, A. Ashikhmin, T.L. Marzetta, S. Vishwanath, Pilot contamination and precoding in multi-cell TDD systems, *IEEE Trans. Wirel. Commun.* 10 (2011) 2640–2651.
- [73] E. Björnson, L. Sanguinetti, J. Hoydis, M. Debbah, Optimal design of energy-efficient multi-user MIMO systems: is massive MIMO the answer? *IEEE Trans. Wirel. Commun.* 14 (2015) 3059–3075.
- [74] S. Jin, X. Wang, Z. Li, K.-K. Wong, Y. Huang, X. Tang, On massive MIMO zero-forcing transceiver using time-shifted pilots, *IEEE Trans. Veh. Technol.* 65 (2016) 59–74.
- [75] C.K. Wen, S. Jin, K.K. Wong, J.C. Chen, P. Ting, Channel estimation for massive MIMO using Gaussian-mixture Bayesian learning, *IEEE Trans. Wirel. Commun.* 14 (2015) 1356–1368.
- [76] J. Zhang, B. Zhang, S. Chen, X. Mu, M. El-Hajjar, L. Hanzo, Pilot contamination elimination for large-scale multiple-antenna aided OFDM systems, *IEEE J. Sel. Top. Signal Process.* 8 (2014) 759–772.
- [77] X. Guo, S. Chen, J. Zhang, X. Mu, L. Hanzo, Optimal pilot design for pilot contamination elimination/reduction in large-scale multiple-antenna aided OFDM systems, *IEEE Trans. Wirel. Commun.* 15 (11) (2016) 7229–7243.
- [78] H. Huh, A.M. Tulino, G. Caire, Network MIMO with linear zero-forcing beamforming: large system analysis, impact of channel estimation, and reduced-complexity scheduling, *IEEE Trans. Inf. Theory* 58 (2012) 2911–2934.
- [79] A. Ashikhmin, T. Marzetta, Pilot contamination precoding in multi-cell large scale antenna systems, in: Proc. IEEE Int. Symp. Inf. Theory (ISIT), Jul., 2012, pp. 1137–1141.

- [80] M. Filippou, D. Gesbert, H. Yin, Decontaminating pilots in cognitive massive MIMO networks, in: Proc. Int. Symp. Wirel. Commun. Syst. (ISWCS), Paris, France, 2012, pp. 816–820.
- [81] H. Yin, D. Gesbert, M. Filippou, Y. Liu, A coordinated approach to channel estimation in large-scale multiple-antenna systems, IEEE J. Sel. Areas Commun. 31 (2013) 264–273.
- [82] R.R. Möller, L. Cottatellucci, M. Vehkaperä, Blind pilot decontamination, IEEE J. Sel. Top. Signal Process. 8 (2014) 773–786.
- [83] L. Shen, Y.D. Yao, H. Wang, H. Wang, Blind decoding based on independent component analysis for a massive MIMO uplink system in microcell Rician/Rayleigh fading channels, IEEE Trans. Veh. Technol. 65 (2016) 8322–8330.
- [84] H.Q. Ngo, Massive MIMO: Fundamentals and System Designs (Ph.D. Thesis), Linköping University Electronic Press, 2015.
- [85] C. Desset, B. Debaillie, F. Louagie, Modeling the hardware power consumption of large scale antenna systems, in: Proc. IEEE Online Conf. Green Commun., 2014.
- [86] M. Di Renzo, H. Haas, A. Ghayeb, S. Sugiura, L. Hanzo, Spatial modulation for generalized MIMO: challenges, opportunities and implementation, Proc. IEEE 102 (2014) 56–103.
- [87] T. Lakshmi Narasimhan, P. Raviteja, A. Chockalingam, Generalized spatial modulation in large-scale multiuser MIMO systems, IEEE Trans. Wirel. Commun. 14 (2015) 3764–3779.
- [88] T. Lakshmi Narasimhan, A. Chockalingam, On the capacity and performance of generalized spatial modulation, IEEE Commun. Lett. 20 (2016) 252–255.
- [89] R.G. Parker, R.L. Rardin, Discrete Optimization, Academic Press, San Diego, CA, 1988.
- [90] Y. Zeng, R. Zhang, Z.N. Chen, Electromagnetic lens-focusing antenna enabled massive MIMO: performance improvement and cost reduction, IEEE J. Sel. Areas Commun. 32 (2014) 1194–1206.
- [91] T. Kwon, Y.G. Lim, B.W. Min, C.B. Chae, RF lens-embedded massive MIMO systems: fabrication issues and codebook design, IEEE Trans. Microw. Theory Tech. 64 (2016) 2256–2271.
- [92] T. Kim, J. Park, J.-Y. Seol, S. Jeong, J. Cho, W. Roh, Tens of Gbps support with mmWave beamforming systems for next generation communications, in: Proc. IEEE Global Commun. Conf. (GLOBECOM), 2013, pp. 3685–3690.
- [93] S. Han, I. Chih-Lin, Z. Xu, C. Rowell, Large-scale antenna systems with hybrid analog and digital beamforming for millimeter wave 5G, IEEE Commun. Mag. 53 (2015) 186–194.
- [94] R.W. Heath, N. González-Prelcic, S. Rangan, W. Roh, A.M. Sayeed, An overview of signal processing techniques for millimeter wave MIMO systems, IEEE J. Sel. Top. Signal Process. 10 (2016) 436–453.
- [95] H. Huh, G. Caire, H.C. Papadopoulos, S.A. Ramprashad, Achieving “massive MIMO” spectral efficiency with a not-so-large number of antennas, IEEE Trans. Wirel. Commun. 11 (2012) 3226–3239.

# Recent advances in network beamforming

# 9

**Shahram ShahbazPanahi\***, **Yindi Jing†**

*University of Ontario Institute of Technology, Oshawa, ON, Canada\** *University of Alberta, Edmonton, AB, Canada†*

## 9.1 INTRODUCTION

Beamforming is a powerful technique which has been widely used in signal processing, radar and sonar, biomedical, and particularly in communications. In applications of beamforming in communications systems, the basic idea is to optimally process signals received over different antennas, or the signals which are to be transmitted over different paths, by adjusting the signal amplitudes and phases, to form a strong beam toward the direction of interest, and at the same time, to avoid receiving or creating interference. Both receive beamforming and transmit beamforming have been extensively studied in the literature leading to numerous innovative and interesting beamforming schemes. In the past decade, with the booming research in cooperative relay communication, a new type of beamforming, namely network beamforming or relay beamforming, has appeared in the literature. A network beamformer is implemented neither at the transmitter nor at the receiver but at the intermediate relays to adaptively process signals transmitted via different relay paths. In fact, one can view a network beamformer performing simultaneously the task of a receive beamformer and that of a transmit beamformer. Network beamforming is an integrated combination of receive beamforming applied to the signals received at different relays from one or more sources and transmit beamforming applied to the signals sent from different relays to the corresponding destinations. A prominent distinction of network beamforming compared to conventional transmit and receive beamforming is that network beamforming is performed somewhere between the two ends of a channel which connects a source to a destination, instead of at either end. Thus, a virtual kind of channel design of the communication systems is realized via network beamforming. Successful network beamforming solutions have shown to provide large data-rates and/or improve link reliability. These advantages offered by network beamforming have presented new challenges to the research community. Research in this area has led to new analytical and optimization tools, which can be beneficial in solving communication problems in general.

In this book chapter, a systematic overview of various network beamforming approaches and solutions are presented for multi-relay cooperative networks with a

wide variety of network configurations, including one-way and two-way communications, single-user and multi-user cases, flat-fading and frequency-selective channel models, and networks with perfect and partial channel state information (CSI). Section 9.2 presents a general model for the end-to-end channel in relay networks and introduces common notation used throughout the chapter. Section 9.3 focuses on network beamforming for one-way relay networks, where both flat-fading channels and frequency-selective channels are considered separately for single-user and multi-user networks. Section 9.4 surveys recent results on network beamforming for two-way relay networks, covering synchronous networks, asynchronous networks, and networks with frequency-selective channels. It is worth mentioning that even though the goal of this chapter is to be as comprehensive and as cohesive as possible, achieving these two goals simultaneously turned out to be almost impossible, given the large volume of the work appeared in the literature on network beamforming, our limited knowledge, and the page limit. As such, many important works may have been left out.

*Notations:* Throughout the chapter, vectors and matrices are represented by bold lower and upper letters, respectively, while calligraphic letters (such as  $\mathcal{U}$  and  $\mathcal{W}$ ) stand for sets. Matrix transpose and Hermitian operators are denoted as  $(\cdot)^T$  and  $(\cdot)^H$ , respectively, while  $(\cdot)^*$  stands for complex conjugate. Schur-Hadamard (element-wise) and Kronecker matrix products are represented by  $\odot$  and  $\otimes$ , respectively. Amplitude and phase of a complex number are denoted as  $|\cdot|$  and  $\angle \cdot$ , respectively. The  $\ell_2$  norm of a vector is represented as  $\|\cdot\|$ , while  $\|\cdot\|_F$  stands for the Frobenius norm of a matrix.  $\text{diag}(\mathbf{a})$  stands for a diagonal matrix whose diagonal entries are the elements of vector  $\mathbf{a}$ . The discrete-time convolution is represented by  $\star$ . The notation  $\mathbf{a} \succeq 0$  indicates that the elements of the vector  $\mathbf{a}$  are nonnegative, while  $\mathbf{A} \succeq (\succ)0$  means matrix  $\mathbf{A}$  is positive semidefinite (positive definite). The statistical expectation is denoted by  $E\{\cdot\}$ .  $\mathcal{P}\{\cdot\}$  stands for the normalized principle eigenvector of a matrix, and  $\lambda_{\max}\{\cdot\}$  represents the principle eigenvalue of a matrix.

## 9.2 END-TO-END CHANNEL MODELING

Considered in this chapter is a network consisting of multiple (say  $N_p$ ) *single-antenna* transceiver (or transmitter-receiver) pairs. The two nodes in each pair wish to communicate with each other (in one direction or in both directions) with the help of multiple (say  $N_r$ ) half-duplex *single-antenna*<sup>1</sup> relays. Each relay produces and transmits a linearly transformed version of its received signal, which is obtained by passing this signal through a finite impulse response (FIR) filter. In a static scenario, the linear time invariant (LTI) end-to-end channel from Transceiver (or Transmitter)  $p$  to Transceiver (or Receiver)  $q$  can be represented using the channel impulse response (CIR), denoted as  $h_{pq}[n]$ , and is given by

$$h_{pq}[n] = \sum_{r=1}^{N_r} \check{f}_{pr}[n] \star w_r^*[n] \star \check{g}_{rq}[n], \quad (9.1)$$

---

<sup>1</sup>In this book chapter, the focus is on single-antenna relaying.

where  $\star$  stands for discrete-time convolution,  $*$  represents complex conjugate,  $w_r[n]$  is the conjugate of the impulse response of the FIR filter used in the  $r$ th relay,  $\check{f}_{pr}[n]$  is the CIR of the link from Transceiver (Transmitter)  $p$  to the  $r$ th relay,  $\check{g}_{rq}[n]$  is the CIR of the link from the  $r$ th relay to Transceiver (Receiver)  $q$ . Note that the discrete-time channel model in Eq. (9.1) assumes the time granularity of one symbol period. In other words, this CIR represents the end-to-end channel from the input of the pulse shaping filter at the transmitter front-end up to the output of receiver filter (matched to the transmitted pulse) filter at the receiver.

The relaying protocol, where each relays uses an FIR filter to process the relay received signals, is referred to as filter-and-forward (FF) relaying scheme [1–5]. The FF scheme was first used in one-way relay channels with frequency-selective source-relay and destination-relay links as a means to equalize the end-to-end channel in a distributed manner [1, 3, 4]. The FF scheme was later used in the context of two-way relay networks with frequency-selective transceiver-relay links [2, 5].

Note that in the case of amplify-and-forward (AF) relaying scheme, the FIR filters  $\{w_r[\cdot]\}_{r=1}^{N_r}$  have only one tap. In this case, each relay uses a single complex coefficient, often called relay beamforming weight, to amplify and to adjust the phase of the relay received signal, and then, retransmits the so-obtained signal. In the AF scheme, the complex beamforming weight of the  $r$ th relay is denoted as  $w_r$  and the vector of the beamforming weights of all relays is denoted as  $\mathbf{w} \triangleq [w_1 \ w_2 \ \dots \ w_{N_r}]^T$ .

One aspect of the channel model in Eq. (9.1) is that unlike traditional channel models (where the end-to-end channel is fixed and is out of the control of system designer), the end-to-end CIR in Eq. (9.1) is somewhat under the control of the system designer, through designing the relay FIR filters. As a result, the problem of designing transmit strategies is intertwined with *channel design*—a concept which does not appear when dealing with traditional multi-path channel models (which we herein refer to as passive channels). A channel whose impulse response can be designed in some optimal sense, is herein referred to as an active channel. A relay channel can indeed be viewed as an active channel. Interested readers are referred to [6–9] for more details on sum-rate-optimal design of active channels.

In the rest of this chapter, the following notation is used: in a single-carrier scheme, the transmit power of the  $q$ th source is denoted as  $P_q$ , while in a multicarrier system, the transmit power of the  $q$ th transmitter over the  $i$ th subcarrier is represented as  $P_{iq}$ .

### 9.3 ONE-WAY NETWORK BEAMFORMING

This section focuses on the concept, formulation, design, and results related to network beamforming for one-way relay-assisted communication, where the information transmissions are established from one or more sources to the corresponding destinations through the relays. As a result, the channels from the destination(s) to the relays and from the relays to the source(s) are not relevant. First, the focus will be on synchronous networks with frequency-flat source-relay and relay-destination links.

The case with frequency-selective source-relay and relay-destination links will be discussed in the subsequent subsection.

In this section, without loss of generality, all noises are assumed to be independent circularly symmetric complex Gaussian random variables with zero-mean and unit-variance. For non-unit-variance noise, results presented in this section can be easily adopted by properly scaling the power parameters.

### 9.3.1 NETWORKS WITH FREQUENCY-FLAT CHANNELS

In networks with frequency-flat source-relay and relay-destination links, each channel can be modeled with a complex coefficient in the baseband representation. Based on the notation described in [Section 9.2](#), the channel from Source  $p$  to Relay  $r$  is denoted as  $f_{S_p,r}$  and the channel from Relay  $r$  to Destination  $q$  is represented as  $g_{r,D_q}$ . The transmit power of Source  $p$  is denoted as  $P_p$  and its corresponding power limit is denoted as  $P_p^{\max}$ . The transmit power of Relay  $r$  is denoted as  $Q_r$  and its corresponding power limit is denoted as  $Q_r^{\max}$ .

For synchronous networks with frequency-flat channels, as there is no inter-symbol-interference (ISI), the transmission of each symbol can be considered separately. In relay beamforming, one complex coefficient is used at each relay to amplify and phase-adjust the relay received signal, thereby allowing the relays to collectively form desired beams toward the destinations. Recall from [Section 9.2](#) that the beamforming coefficient of Relay  $r$  is denoted as  $w_r$  and the  $N_r \times 1$  relay beamforming vector containing the coefficients of all relays is represented as  $\mathbf{w} = [w_1 \ w_2 \ \dots \ w_{N_r}]^T$ .

#### 9.3.1.1 Single-user networks

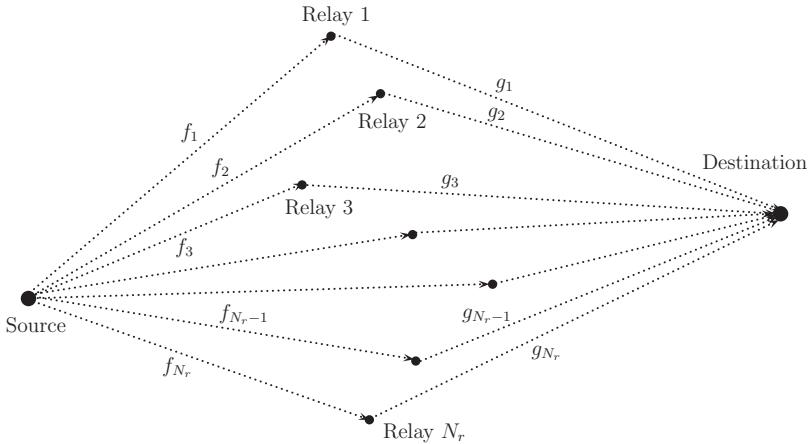
Considered in this subsection is a relay network with one source, one destination, and  $N_r$  single-antenna relays (which are indexed as  $1, 2, \dots, N_r$ ), as depicted in [Fig. 9.1](#). Since only one source and one destination are present in the network, the notation is simplified by denoting the channel from the source to Relay  $r$  as  $f_r$  and the channel from Relay  $r$  to the destination as  $g_r$ . The vector  $\mathbf{f} \triangleq [f_1 \ f_2 \ \dots \ f_{N_r}]^T$  is the channel vector from the source to all relays and  $\mathbf{g} \triangleq [g_1 \ g_2 \ \dots \ g_{N_r}]^T$  is the channel vector from all relays to the destination. The transmit power of the source is denoted as  $P$ .

In a two-step relay-assisted communication protocol, the source transmits the signal  $\sqrt{P}s$ , where  $s$  is the unit-power information symbol (i.e.,  $E\{|s|^2\} = 1$ ). The received signal vector at the relays is given by

$$\mathbf{x} = \sqrt{P}\mathbf{f}s + \mathbf{v}, \quad (9.2)$$

where  $\mathbf{v}$  is the vector of the relay noises. In the second step, the  $r$ th relay multiplies its received signal by  $w_r^*$  and transmits the so-obtained signal toward the destination. The  $N_r \times 1$  vector of the relay transmitted signals is denoted as  $\mathbf{t}$  and is expressed as

$$\mathbf{t} = \mathbf{w}^* \odot \mathbf{x}. \quad (9.3)$$

**FIG. 9.1**

A single-user one-way relay network.

In light of Eqs. (9.2) and (9.3), the signal received at the destination is then written as

$$y = \sqrt{P} \mathbf{w}^H (\mathbf{g} \odot \mathbf{f}) s + (\mathbf{w}^* \odot \mathbf{g}) \mathbf{v} + n, \quad (9.4)$$

where  $n$  is the noise at the destination. Given Eq. (9.4), the end-to-end received SNR at the destination can be expressed as

$$\text{SNR} = \frac{P |\mathbf{w}^H (\mathbf{f} \odot \mathbf{g})|^2}{1 + \|\mathbf{w} \odot \mathbf{g}\|^2}. \quad (9.5)$$

The transmit power used at Relay  $r$  can be calculated, using Eqs. (9.2) and (9.3), as

$$Q_r = |w_r|^2 (1 + P |f_r|^2). \quad (9.6)$$

Based on Eq. (9.6), the total transmit power consumed by all relays is thus expressed as

$$Q_T = P \|\mathbf{w} \odot \mathbf{b}\|^2, \quad (9.7)$$

where the  $N_r \times 1$  vector  $\mathbf{b}$  is defined as

$$\mathbf{b} \triangleq \begin{bmatrix} \sqrt{\frac{1}{P} + |f_1|^2} & \sqrt{\frac{1}{P} + |f_2|^2} & \dots & \sqrt{\frac{1}{P} + |f_{N_r}|^2} \end{bmatrix}^T.$$

The goal of relay beamforming design is to find the optimal  $\mathbf{w}$  with respect to a performance measure under certain resource or quality-of-service (QoS) constraints.

A significant volume of literature has presented extensive results on relay beamforming for single-user frequency-flat synchronous relay networks. In the rest of this subsection, we review some of the exemplary results.

### 9.3.1.2 SNR-maximization with perfect CSI

For single-user networks with a single source-destination pair, the end-to-end received SNR maximization is equivalent to the rate maximization or the optimization of a few other performance metrics, such as outage probability and bit error rate (BER). In this part, perfect CSI is assumed. The authors of [10] study the problem of relay beamforming design based on the maximization of the end-to-end received SNR under a total relay power constraint. This problem can be written as

$$\max_{\mathbf{w}} \text{SNR}, \text{ subject to } Q_T \leq Q_T^{\max},$$

which can be specified, using Eqs. (9.5) and (9.7), as

$$\max_{\mathbf{w}} \frac{P|\mathbf{w}^H(\mathbf{f} \odot \mathbf{g})|^2}{1 + \|\mathbf{w} \odot \mathbf{g}\|^2}, \text{ subject to } P \|\mathbf{w} \odot \mathbf{b}\|^2 \leq Q_T^{\max}. \quad (9.8)$$

As proven in [10], the SNR maximization problem (9.8) is amenable to a neat closed-form solution as summarized below.

#### RESULT 9.1

For single-user multi-relay networks, the optimal relay beamforming solution which maximizes the end-to-end SNR under a total relay power budget of  $Q_T^{\max}$ , is given, for the  $r$ th relay, as

$$|w_r| = C_{\text{sum}} \frac{|f_r g_r|}{1 + P|f_r|^2 + Q_T^{\max}|g_r|^2} \sqrt{Q_T^{\max}}, \quad \angle w_r^o = \angle f_r + \angle g_r,$$

where

$$C_{\text{sum}} \triangleq \left[ \sum_{r'=1}^{N_r} \frac{|f_{r'} g_{r'}|^2 (1 + P|f_{r'}|^2)}{(1 + P|f_{r'}|^2 + Q_T^{\max}|g_{r'}|^2)^2} \right]^{-1/2}.$$

With this optimal relay beamforming solution, the end-to-end received SNR can be calculated as

$$\text{SNR}^o = \sum_{r=1}^{N_r} \frac{P Q_T^{\max} |f_r g_r|^2}{1 + P|f_r|^2 + Q_T^{\max}|g_r|^2}. \quad (9.9)$$

**Result 9.1** provides the following insights into the relay beamforming design. First, due to the phase alignment, each relay should perfectly cancel the sum of the phases of the complex coefficients which represent the links between that relay and the two end-nodes, thereby allowing the transmissions through different relay paths to add up coherently. Second, with a total relay power constraint, the power allocated to Relay  $r$  is proportional to  $\frac{|f_r g_r|^2}{1 + P|f_r|^2 + Q_T^{\max}|g_r|^2}$ , indicating that the two communication hops affect the power allocation symmetrically. Third, the optimal beamforming weight of each relay only depends on the local channel coefficients of that specific relay, except for the factor  $C_{\text{sum}}$  which is common to all relays. The beamforming can thus be implemented in a distributed manner in the sense that each relay needs only the knowledge of its own channel coefficients.

The common factor  $C_{\text{sum}}$  can be broadcasted from the destination to all relays. Finally, the solution in [Result 9.1](#) shows that the end-to-end received SNR with the optimal beamforming is the sum of the end-to-end SNR along every relay path. Thus, the optimal beamforming has the same effect as the maximum-ratio combining technique in multi-input single-output (MISO) systems, and can be seen as the extension of this technique to dual-hop relay networks.

For the same network described previously, the investigation in [11] considers the SNR maximization approach to design a network beamformer under the condition that each individual relay has its own power constraint. The problem of relay beamforming design via maximizing the end-to-end SNR under the individual relay power constraint can be written as

$$\max_{\mathbf{w}} \text{SNR}, \text{ subject to } Q_r \leq Q_r^{\max}, \text{ for } r = 1, 2, \dots, N_r,$$

which can be specified, using Eqs. (9.5) and (9.6), as

$$\max_{\mathbf{w}} \frac{P|\mathbf{w}^H(\mathbf{f} \odot \mathbf{g})|^2}{1 + \|\mathbf{w} \odot \mathbf{g}\|^2}, \text{ subject to } |w_r|^2 (1 + P|f_r|^2) \leq Q_r^{\max}, \text{ for } r = 1, 2, \dots, N_r. \quad (9.10)$$

Unlike the optimization problem (9.8) which has only one constraint, the SNR maximization problem (9.10) has  $N_r$  constraints. Furthermore, the feasible region in the optimization problem (9.10) is a hypercube, which has a non-smooth surface. However, the feasible region in the optimization problem (9.10) is a hypersphere, which has a smooth surface. As such, solving the optimization problem (9.10) is significantly more challenging than solving the optimization problem (9.8). Nevertheless, the sophisticated derivations of [11] present a closed-form solution to the SNR maximization problem under individual per-relay power constraints. This solution is summarized below.

## RESULT 9.2

Define

$$\phi_r = \frac{\sqrt{1 + P|f_r|^2} \cdot |f_r|}{\sqrt{Q_r^{\max}} \cdot |g_r|}, \text{ for } r = 1, 2, \dots, N_r.$$

Let  $(\xi_1, \xi_2, \dots, \xi_{N_r})$  be the permutation of  $(1, 2, \dots, N_r)$  such that  $\phi_{\xi_1} > \phi_{\xi_2} > \dots > \phi_{\xi_{N_r}} > \phi_{\xi_{N_r+1}} \triangleq 0$ . Define, for  $r = 1, 2, \dots, N_r$ ,

$$\lambda_r = \frac{1 + \sum_{m=1}^r \frac{Q_{\xi_m}^{\max} |g_{\xi_m}|^2}{1 + P|f_{\xi_m}|^2}}{\sum_{m=1}^r \frac{\sqrt{Q_{\xi_m}^{\max} \cdot |f_{\xi_m} g_{\xi_m}|}}{\sqrt{1 + P|f_{\xi_m}|^2}}}.$$

Let  $r_0$  be the smallest  $r$  such that  $\lambda_r < \phi_{\xi_{r+1}}^{-1}$ .

For single-user multi-relay networks, the optimal relay beamforming design which maximizes the end-to-end received SNR under the individual relay power constraints,  $Q_1^{\max}, \dots, Q_{N_r}^{\max}$ , is as follows.

*Continued*

**RESULT 9.2—CONT'D**

$$|w_r| = \begin{cases} \frac{\sqrt{Q_r^{\max}}}{\sqrt{1 + P|f_r|^2}} & r = \xi_1, \dots, \xi_{r_0} \\ \lambda_{r_0} \frac{|f_r|}{|g_r|} & r = \xi_{r_0+1}, \dots, \xi_{N_r} \end{cases}, \quad \angle w_r = \angle f_r + \angle g_r.$$

With this optimal relay beamforming solution, the transmit power of Relay  $r$  is

$$Q_r = \begin{cases} Q_r^{\max} & r = \xi_1, \dots, \xi_{r_0} \\ \lambda_{r_0} \phi_r^2 Q_r^{\max} & r = \xi_{r_0+1}, \dots, \xi_{N_r} \end{cases}.$$

The optimal relay beamforming solution in [Result 9.2](#), although not as tidy as the one presented in [Result 9.1](#) for the case of the total relay power constraint, is in closed form since the exact value can be obtained within a finite number of operations. The complexity in finding the solution is  $\mathcal{O}(M \log M)$ , where the step with the highest complexity is the ordering of  $\phi_r$ 's. [Result 9.2](#) shows that the phase alignment at the relays is the same as that in [Result 9.1](#) and that the optimal power used at a relay can be any value between 0 and its maximum allowable power consumption. At the optimum, the transmit power of each relay depends not only on the channel coefficients of that relay but also on the channel coefficients of all other relays. As a result, the optimal solution is not an on-or-off one, not a decoupled one, and, in general, not even a differentiable function of the channel coefficients. [Result 9.2](#) has stimulated research on power control for relay networks. In addition, it can be seen from [Result 9.2](#) that the  $r_0$  relays with the highest  $\phi_r$  values use their maximum power, while the rest of the relays use fractions of their powers that are proportional to  $\lambda_{r_0}^2 |f_r/g_r|^2 (1 + |f_r|^2 P)$ . Since the power proportion of each relay depends only on the channels of that relay except for the coefficient  $\lambda_{r_0}$  which is common to all relays, the optimal relay beamforming solution can be implemented in a distributed fashion where each relay requires only its local CSI along with a small amount of information broadcasted by the destination. Two distributed implantations are proposed in [11].

In [12], for the same network, the SNR maximization problem was studied under both total and individual relay power constraints. It was shown that the optimal solution can be found numerically using second-order conic programming. This solution is amenable to a distributed implementation.

### **9.3.1.3 SNR-per-unit-power maximization**

In [13], a new metric, namely SNR-per-unit-power, is proposed for designing relay beamformers. The SNR-per-unit-power is indeed proposed as an efficiency measure for relay networks. In the literature, there has been a significant volume of research on two most popular efficiency measures: spectral efficiency and energy efficiency. Spectral efficiency is the achievable bit-rate and its maximization guarantees the highest

amount of information flow for given transmit power. This measure however does not consider how efficient the power is used in achieving the maximum. Energy efficiency, defined as the number of transmitted bits per unit energy or power, is a more natural efficiency measure. But for most systems, the maximum of energy efficiency is achieved when the transmit power approaches 0. Hence, energy-efficiency-based designs trap the system in low SNR regime and deliver low service quality. Inspired by the limitations of these efficiency measures, the SNR-per-unit-power (or power-normalized SNR or power efficiency [14]) for single-user network is defined as

$$\eta \triangleq \frac{\text{SNR}}{P_T}, \quad (9.11)$$

where SNR is the end-to-end received SNR and  $P_T$  is the total power consumed in the network. This new efficiency measure does not trap the system in low power regime.

Assuming perfect CSI, the study in [13] investigates thoroughly the maximization of the SNR-per-unit-power for single-relay networks, derives the performance of the so-obtained optimal design, and compares this design with other efficiency-optimal designs. Also studied in [13] are relay beamforming designs for multi-relay networks under both total relay power constraint and individual relay power constraints. In the sequel, such multi-relay beamforming designs are presented.

With the aforementioned network model, the total power consumed in the network is given as

$$P_T = P + P \|\mathbf{w} \odot \mathbf{b}\|^2. \quad (9.12)$$

From Eqs. (9.5) and (9.12), the SNR-per-unit-power of the relay network can be calculated as

$$\eta = \frac{|\mathbf{w}^H(\mathbf{f} \odot \mathbf{g})|^2}{(1 + \|\mathbf{w} \odot \mathbf{g}\|^2)(1 + \|\mathbf{w} \odot \mathbf{b}\|^2)}. \quad (9.13)$$

The SNR-per-unit-power maximization problem under a total relay power budget of  $Q_T^{\max}$  can be formulated as

$$\min_{\mathbf{w}} \frac{|\mathbf{w}^H(\mathbf{f} \odot \mathbf{g})|^2}{(1 + \|\mathbf{w} \odot \mathbf{g}\|^2)(1 + \|\mathbf{w} \odot \mathbf{b}\|^2)}, \text{ subject to } P \|\mathbf{w} \odot \mathbf{b}\|^2 \leq Q_T^{\max} \quad (9.14)$$

and the corresponding problem under individual per relay power constraints amounts to the following one:

$$\min_{\mathbf{w}} \frac{|\mathbf{w}^H(\mathbf{f} \odot \mathbf{g})|^2}{(1 + \|\mathbf{w} \odot \mathbf{g}\|^2)(1 + \|\mathbf{w} \odot \mathbf{b}\|^2)}, \text{ subject to } |w_r|^2 (1 + P|f_r|^2) \leq Q_r^{\max} \text{ for } r = 1, 2, \dots, N_r. \quad (9.15)$$

One can easily show that at the optimum of the optimization problem (9.14) and at the optimum of the optimization problem (9.15), each relay should perfectly cancel the total-phase of its channel coefficients, i.e., the optimal solution for the phases of  $\mathbf{w}$  is  $\angle w_i = \angle f_i + \angle g_i$ . The main question is how to find the amplitudes of the relay beamforming weights. The major results of [13] are summarized in [Result 9.3](#).

### RESULT 9.3

For single-user multi-relay networks, the optimal relay beamforming design which maximizes the SNR-per-unit-power under the total relay power constraint  $Q_T^{\max}$  can be reduced (by using [Result 9.1](#)) to the following one-dimensional problem:

$$\max_{0 \leq Q_T \leq Q_T^{\max}} \sum_{r=1}^R \frac{PQ_T |f_r|^2 |g_r|^2}{(P|f_r|^2 + Q_T |g_r|^2 + 1)(P + Q_T)}. \quad (9.16)$$

The objective function the optimization problem (9.16) is a semistrictly quasiconcave function and has only one maximum for  $Q_T > 0$ . Thus, the gradient-ascent technique can be used to find the optimal solution.

For single-user multi-relay networks, the optimal relay beamforming design which maximizes the SNR-per-unit-power under the individual relay power constraints,  $Q_1^{\max}, \dots, Q_N^{\max}$ , can be found by using a combination of sequential quadratic programming and scatter search. A low-complexity suboptimal solution is proposed in [13] via constraint relaxation and projection, where the optimal solution without the power constraints is first found via one-dimensional gradient-ascent method; then the result is projected onto the feasible region defined by the individual power constraints.

#### 9.3.1.4 Partial CSI

The results of [10–13], summarized in the previous subsection assumes perfect CSI. However, in practice, due to user mobility, overhead, and resource constraints, relays may only have access to partial CSI. Two common partial CSI models are the quantized CSI and the statistical CSI.

In [15], the authors investigate the relay beamforming design problem for the single-user multi-relay network with quantized CSI at the relays. It is assumed in [15] that only the receiver has perfect CSI and the relays can only obtain quantized CSI information from receiver feedback and use that for beamforming design. In the approach presented in [15], a codebook of beamforming vectors is used at the relays. For each transmission, a beamforming vector is selected based on an index from the receiver feedback. The relay beamforming problem is thus the joint optimization of the quantizer at the receiver (to find the optimal beamforming vector index based on the CSI) and the quantization codebook for a fixed number of feedback bits. The study in [15] uses a BER minimization approach to solve this design problem. Given a beamforming codebook, the BER-minimizing quantizer chooses the beamforming vector in the codebook that maximizes the received SNR. For the codebook design, the generalized Lloyd algorithm is used. Other than the relay beamforming design, several analytical performance results are also derived in [15]. First, relay selection can be seen as a special case of the relay beamforming design where the elements of the codebook are standard basis vectors. It is proved rigorously in [15] that relay selection can achieve the optimal diversity order. Second, the authors of [15] quantitatively analyze the average SNR loss and the capacity loss of the beamforming design under quantized CSI compared to perfect CSI case, thereby showing that these losses decay at least exponentially as the number of feedback bits grows. Moreover, a tight approximation on the BER is derived.

In [16], relay beamforming is investigated for the single-user multi-relay network with second-order statistical information of the channels as the partial CSI. Two

beamforming approaches are proposed in [16]: a total relay transmit power minimization approach subject to SNR requirement and an SNR maximization approach subject to total or individual relay power constraints. Note that since only statistical CSI is assumed to be available, in deriving the average relay transmit power, the average received noise power, and the SNR, one has to take expectation over all noises and over all channels, while in previous work where perfect instantaneous CSI is available, the average is taken over the noises only.

Assuming statistical CSI, the relay beamforming design which minimizes the total relay transmit power subject to SNR requirement of at least  $\gamma$  can be formulated as

$$\min_{\mathbf{w}} Q_T, \text{ subject to } \text{SNR} \geq \gamma. \quad (9.17)$$

Using Eqs. (9.5) and (9.7) and assuming second-order CSI, one can write the optimization problem (9.17) as

$$\min_{\mathbf{w}} P \|\mathbf{w} \odot \tilde{\mathbf{b}}\|^2, \text{ subject to } \frac{P\mathbf{w}^H \mathbf{R}\mathbf{w}}{1 + \mathbf{w}^H \mathbf{Q}\mathbf{w}} \geq \gamma, \quad (9.18)$$

where  $\mathbf{R} \triangleq \mathbb{E}\{(\mathbf{f} \odot \mathbf{g})(\mathbf{f} \odot \mathbf{g})^H\}$  is the  $N_r \times N_r$  correlation matrix of the end-to-end channel vector  $\mathbf{f} \odot \mathbf{g}$ , the  $\mathbf{Q} \triangleq \mathbb{E}\{\mathbf{g}\mathbf{g}^H\}$  is the  $N_r \times N_r$  correlation matrix of the relay-destination channel vector  $\mathbf{g}$ , and the vector  $\tilde{\mathbf{b}}$  is defined as

$$\tilde{\mathbf{b}} \triangleq \left[ \sqrt{\frac{1}{P} + \mathbb{E}[|f_1|^2]} \quad \dots \quad \sqrt{\frac{1}{P} + \mathbb{E}[|f_{N_r}|^2]} \right]^T.$$

Using Lagrange multiplier method, one can show that the optimization problem (9.18) has a closed-form solution, as summarized below.

#### RESULT 9.4

For the single-user multi-relay network with second-order statistical CSI, the optimal relay beamforming vector that minimizes the total relay power for a given feasible SNR lower limit  $\gamma$  is given by

$$\mathbf{w}^o = \sqrt{\frac{\gamma}{P\lambda_{\max}\{[\text{diag}\{\tilde{\mathbf{b}}\}]^{-1}(\mathbf{R} - \frac{\gamma}{P}\mathbf{Q})[\text{diag}\{\tilde{\mathbf{b}}\}]^{-1}\}} \text{diag}\{\tilde{\mathbf{b}}\}^{-1} \frac{\mathbf{u}^*}{\|\mathbf{u}\|}}.$$

Here,  $\mathbf{Q}$  is the correlation matrix of the vector  $\mathbf{g}$ ,  $\mathbf{R}$  is the correlation matrix of the vector  $\mathbf{f} \odot \mathbf{g}$ , and  $\mathbf{u} = \mathcal{P}\{[\text{diag}\{\tilde{\mathbf{b}}\}]^{-1}(\mathbf{R} - \frac{\gamma}{P}\mathbf{Q})[\text{diag}\{\tilde{\mathbf{b}}\}]^{-1}\}$ , where  $\mathcal{P}\{\cdot\}$  stands for the normalized principle eigenvector of a matrix, and  $\lambda_{\max}\{\cdot\}$  represents the principle eigenvalue of a matrix. The corresponding minimum total relay transmit power is obtained as

$$Q_T^o = \frac{\gamma}{\lambda_{\max}\{[\text{diag}\{\tilde{\mathbf{b}}\}]^{-1}(\mathbf{R} - \frac{\gamma}{P}\mathbf{Q})[\text{diag}\{\tilde{\mathbf{b}}\}]^{-1}\}}.$$

The relay beamforming designs which maximize the end-to-end SNR under a total relay power constraint and under individual relay power constraints can be formulated, respectively, as

$$\max_{\mathbf{w}} \frac{P\mathbf{w}^H \mathbf{R}\mathbf{w}}{1 + \mathbf{w}^H \mathbf{Q}\mathbf{w}}, \text{ subject to } P \|\mathbf{w} \odot \tilde{\mathbf{b}}\|^2 \leq Q_T^{\max}$$

and

$$\max_{\mathbf{w}} \frac{P\mathbf{w}^H \mathbf{R}\mathbf{w}}{1 + \mathbf{w}^H \mathbf{Q}\mathbf{w}}, \text{ subject to } |w_r|^2 \left(1 + PE[|f_r|^2]\right) \leq Q_r^{\max}, \text{ for } r = 1, 2, \dots, N_r.$$

The main results for these designs are given below.

### RESULT 9.5

For the single-user multi-relay network with second-order statistical CSI, the optimal relay beamforming vector which maximizes the end-to-end SNR for a total relay power limit  $Q_T^{\max}$  is given as

$$\mathbf{w}^o = \sqrt{\frac{Q_T^{\max}}{P}} [\text{diag}\{\tilde{\mathbf{b}}\}]^{-1} \mathcal{P} \left\{ \left( \mathbf{I} + \frac{Q_T^{\max}}{P} [\text{diag}\{\tilde{\mathbf{b}}\}]^{-1} \mathbf{Q} [\text{diag}\{\tilde{\mathbf{b}}\}]^{-1} \right)^{-1} [\text{diag}\{\tilde{\mathbf{b}}\}]^{-1} \mathbf{R} [\text{diag}\{\tilde{\mathbf{b}}\}]^{-1} \right\}^*.$$

The corresponding maximum SNR is then obtained as

$$\text{SNR}^o = Q_T^{\max} \lambda_{\max} \left\{ \left( \mathbf{I} + \frac{Q_T^{\max}}{P} [\text{diag}\{\tilde{\mathbf{b}}\}]^{-1} \mathbf{Q} [\text{diag}\{\tilde{\mathbf{b}}\}]^{-1} \right)^{-1} [\text{diag}\{\tilde{\mathbf{b}}\}]^{-1} \mathbf{R} [\text{diag}\{\tilde{\mathbf{b}}\}]^{-1} \right\}.$$

The SNR maximization under individual relay power constraint problem can be turned into a convex feasibility semidefinite programming (SDP) via semidefinite relaxation. It can be solved using an iterative procedure, where at each step, a convex feasibility problem is solved using an interior point method.

The SNR maximization under the individual relay power constraint problem is further investigated in [17], where it is shown that the semi-definite relaxation of the original problem leads to an SDP problem which always has a rank-one solution. Hence, the original problem can be solved by finding the rank-one solution to this SDP problem. In addition, the SNR maximization under a total source and relay power constraint is studied in [17], where it is shown that the problem can be reduced to a one-dimensional optimization over the source power and can be solved via Newton's method.

#### 9.3.1.5 MSE-minimization and received signal power maximization

The mean-squared-error (MSE) of the received signal and received signal power have also been used as objective functions for relay beamforming design. In [18], the relay beamforming design relies on the minimization of the MSE between the received signal and a scaled version of the transmitted signal with no power constraint. This minimization problem can be formulated as

$$\min_{\mathbf{w}} E\{|\eta s - \mathbf{g}^T \mathbf{t}|^2\}, \quad (9.19)$$

where  $\eta$  is a positive scalar chosen by the designer and  $\mathbf{t}$  is the received signal vector at the relays given in Eq. (9.3). The optimal solution the optimization problem (9.19) is given in [18] as summarized below.

### RESULT 9.6

For the single-user multi-relay network, the optimal relay beamforming weigh of the  $r$ th relay which minimizes the MSE of the received signal and a scaled version of the transmitted signal is given by

$$w_r^0 = \eta \left( \frac{P}{1 + P \|\mathbf{f}\|^2} \right) \frac{f_r g_r}{|g_r|^2}.$$

In addition to the above relay beamforming solution, the optimal equalization at the receiver is derived as well as the MSE, power efficiency, capacity, and pairwise error probability achieved by the optimal design.

To take into consideration the power constraint, the authors of [18] also study two other design approaches which aim to maximize the received signal power at the destination. The problem of maximizing the power of the desired signal component under individual relay power constraints and under a total relay power constraint can be formulated, respectively, as

$$\max_{\mathbf{w}} \quad E\{|\mathbf{g}^T \mathbf{t}|^2\}, \quad \text{subject to} \quad |w_r|^2 \left( 1 + P |f_r|^2 \right) = Q_r^{\max}, \quad \text{for } r = 1, 2, \dots, N_r \quad (9.20)$$

and

$$\max_{\mathbf{w}} \quad E\{|\mathbf{g}^T \mathbf{t}|^2\}, \quad \text{subject to} \quad P \|\mathbf{w} \odot \mathbf{b}\|^2 \leq Q_T^{\max}. \quad (9.21)$$

Note that in the optimization problem (9.20), each relay is constrained to always use its maximum available power. A closed-form solution has been found for the problem (9.20). For the optimization problem (9.21), high SNR approximation can be used to obtain an approximate solution in closed-form. These results are summarized below.

### RESULT 9.7

For the single-user multi-relay network, the optimal relay beamforming vector which minimizes the received signal power at the destination with the separate relay power constraints,  $Q_1^{\max}, \dots, Q_{N_r}^{\max}$  is obtained as

$$w_r^0 = \sqrt{\frac{Q_r^{\max}}{1 + P |f_r|^2}} \frac{f_r g_r}{|f_r| |g_r|}.$$

With the total relay power constraint  $Q_T^{\max}$ , a high SNR approximation of the relay beamforming vector that minimizes the received signal power at the destination is expressed as

$$w_r^0 = \sqrt{\frac{Q_T^{\max}}{\sum_{r'=1}^{N_r} \frac{|f_{r'} g_{r'}|^2}{1 + P |f_{r'}|^2}}} \times \frac{f_r g_r}{1 + P |f_r|^2}.$$

The authors of [19–21] further studied MSE-based relay beamforming designs. The work in [19] focuses on the MSE minimization with respect to both short-term and long-term individual relay power constraints, as well as on the joint design with channel estimation. Here, only results on the short-term power constraint under perfect CSI are reviewed. The corresponding problem can be formulated as

$$\min_{\mathbf{w}} \mathbb{E}\{|s - \mathbf{g}^T \mathbf{t}|^2\}, \text{ subject to } |w_r|^2 \left(1 + P|f_r|^2\right) \leq Q_r^{\max}, \text{ for } r = 1, 2, \dots, N_r. \quad (9.22)$$

In this formulation, all relays have the same power budget of  $Q^{\max}$ . In this work, the beamforming coefficient at each relay is decomposed into two parts, the reception coefficient and the transmission coefficient. The latter is also the preequalizer coefficient. A solution to the optimization problem (9.22) is provided in [19], where the preequalizer coefficient of each relay is represented as a function of the preequalizer coefficients of all other relays. Thus, the implementation of this solution needs numerical iterations. Two closed-form suboptimal solutions, which require only local CSI and no cross-talk among relays or feedback from the receiver, are also proposed.

The first suboptimal relay beamforming solution to optimization problem (9.22) presented in [19] can be expressed as

$$w_r^0 = \frac{P}{(1 + P|f_r|^2)(\lambda_{1,r} + |g_r|^2)} f_r g_r,$$

where  $\lambda_{1,r} = \max\{0, z_r\}$  with  $z_r$  being the largest root of the following polynomial of  $z$ :

$$z^2 + 2|g_r|^2 z + |g_r|^2 \left(|g_r|^2 - \frac{P^2|f_r|^2}{Q_r^{\max}(1 + P|f_r|^2)}\right) = 0.$$

The second suboptimal relay beamforming solution to optimization problem (9.22) presented in [19] can be written as

$$w_r^0 = \frac{P}{(1 + P|f_r|^2)(\lambda_{2,r} + c_r|g_r|^2)} f_r g_r,$$

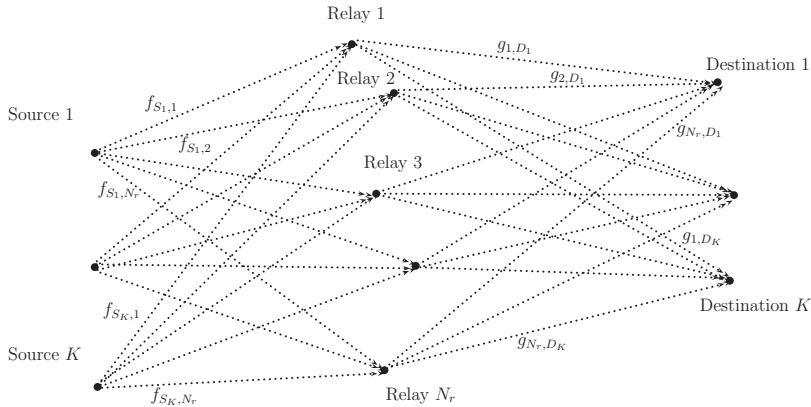
where  $c_r = 1 + \bar{A}_r(M_r - 1)$  with  $\bar{A}_r$  depending on the statistical properties of the channels of Relay  $r$  and  $\lambda_{2,r} = \max\{0, z_r\}$  with  $z_r$  being the largest root of the following polynomial of  $z$ :

$$z^2 + 2c_r|g_r|^2 z + |g_r|^2 \left(c_r|g_r|^2 - \frac{P^2|f_r|^2}{Q_r^{\max}(1 + P|f_r|^2)}\right) = 0.$$

Furthermore, a distributed optimization algorithm is proposed in [20] and a class of adaptive algorithms are proposed in [21] for the relay beamforming problem, both for MSE minimization-based relay beamforming.

### 9.3.1.6 Multi-user networks

This subsection presents the results related to relay beamforming designs for multi-user multi-relay one-way networks with synchronous relays and frequency-flat channels. Consider a network with  $K$  sources (referred to as  $S_1, \dots, S_K$ ),  $K$  destinations

**FIG. 9.2**

A multi-user one-way relay network.

(denoted as  $D_1, \dots, D_K$ ) and  $N_r$  relays. As shown in Fig. 9.2, in this network, Source  $k$  aims to transmit information to Destination  $k$  with the help of the  $N_r$  single-antenna relay nodes. By following the notation system introduced in Section 9.2, let  $f_{S_k,r}$  be the coefficient which represents the channel from  $S_k$  to the  $r$ th relay and  $g_{r,D_k}$  denotes the coefficient of the channel from the  $r$ th relay to  $D_k$ . Let  $\mathbf{f}_k \triangleq [f_{S_k,1} \ f_{S_k,2} \ \dots \ f_{S_k,N_r}]^T$ , and  $\mathbf{g}_k \triangleq [g_{1,D_k} \ g_{2,D_k} \ \dots \ g_{N_r,D_k}]^T$  which are the channel vectors from  $S_k$  to Relay  $r$  and the channel vector from Relay  $r$  to  $D_k$ , respectively. The transmit power of the  $k$ th source is denoted as  $P_k$ .

Having multiple source-destination pairs in the network significantly complicates the relay beamforming design problem since several new challenges have to be addressed in such a network. The first challenge is the design objective. As each source-destination pair has its own QoS level, choosing and formulating the objective function in the multi-user system is far from trivial. Different objectives can lead to totally different designs and performance. Another challenge is the user-interference and relay-interference. The relay beamforming design should be designed not only to achieve coherence aggregation of the relay signals from every source at the corresponding destination but also to effectively attenuate the interference among different source-destination pairs at the same time. Moreover different sources will have conflicting requests for the use of relay power.

### 9.3.1.7 Orthogonal user channels

To generalize the single-user relay network to the multi-user case, one method is to allocate each source-destination pair a distinct orthogonal channel to avoid inter-pair interference [22–24]. However, transmissions of different relays for each pair are nonorthogonal and mingled with each other. In such orthogonal relay-assisted

multiple peer-to-peer communication scenarios, there are  $K$  relay beamforming vectors; denoted as  $\mathbf{w}_1, \dots, \mathbf{w}_K$ ; to be designed, one for each source-destination pair. The  $r$ th entry of the vector  $\mathbf{w}_k$  is denoted as  $w_{S_k,r}$ . The major challenge lies in the allocation of the relay or total power among the sources in helping their transmissions. By following the two-step transmissions as explained earlier, the end-to-end received SNR at Destination  $k$  can be written as

$$\text{SNR}_k = \frac{P_k |\mathbf{w}_k^H(\mathbf{f}_k \odot \mathbf{g}_k)|^2}{1 + \|\mathbf{w}_k \odot \mathbf{g}_k\|^2}. \quad (9.23)$$

The transmit power of Relay  $i$  spent for forwarding the signal of Source  $S_k$  can be expressed as

$$Q_{S_k,r} = |w_{S_k,r}|^2 \left(1 + P_r |f_{S_k,r}|^2\right). \quad (9.24)$$

The total power consumed by all relays for Source  $S_k$  is thus written as

$$Q_{S_k,T} = P_k \|\mathbf{w}_k \odot \mathbf{b}_k\|^2,$$

where

$$\mathbf{b}_k \triangleq \left[ \sqrt{\frac{1}{P_k} + |f_{S_k,1}|^2} \quad \sqrt{\frac{1}{P_k} + |f_{S_k,2}|^2} \quad \dots \quad \sqrt{\frac{1}{P_k} + |f_{S_k,R}|^2} \right]^T. \quad (9.25)$$

The total power used at all relays for all sources is thus

$$Q_T = \sum_{k=1}^K Q_{S_k,T} = \sum_{k=1}^K P_k \|\mathbf{w}_k \odot \mathbf{b}_k\|^2. \quad (9.26)$$

The investigation in [22] studies several approaches to network beamforming designs with orthogonal user channels. The first approach presented in [22] relies on the minimization of the total relay power under  $K$  constraints on the received SNRs at the  $K$  destination. This minimization problem can be cast as

$$\min_{\{\mathbf{w}_k\}_{k=1}^K} \sum_{k=1}^K P_k \|\mathbf{w}_k \odot \mathbf{b}_k\|^2, \text{ subject to } \frac{P_k |\mathbf{w}_k^H(\mathbf{f}_k \odot \mathbf{g}_k)|^2}{1 + \|\mathbf{w}_k \odot \mathbf{g}_k\|^2} \geq \gamma_k, \text{ for } k = 1, 2, \dots, K. \quad (9.27)$$

Based on [Result 9.1](#), the minimization problem (9.27), which is solved over a space of  $KN_r$  complex dimensions, can be transformed to another problem over a space of  $K$  real dimensions, corresponding to the power allocation among different source-destination pairs. Then a numerical fixed-point algorithm can be used to solve the reduced dimension problem [22].

The second approach presented in [22] aims to minimize the total relay power under two sets of constraints. One set of constraints ensures that the received SNRs for all source-destination pairs are above given thresholds and the second set of the constraints limits the powers of individual relays. With Eqs. (9.23), (9.24), and (9.26), this minimization problem can be written as

$$\begin{aligned}
& \min_{\{\mathbf{w}_k\}_{k=1}^K} \sum_{k=1}^K P_k \|\mathbf{w}_k \odot \mathbf{b}_k\|^2, \\
\text{subject to } & \frac{P_k |\mathbf{w}_k^H (\mathbf{f}_k \odot \mathbf{g}_k)|^2}{1 + \|\mathbf{w}_k \odot \mathbf{g}_k\|^2} \geq \gamma_k, \text{ for } k = 1, 2, \dots, K \\
\text{and } & \sum_{k=1}^K |w_{S_k, r}|^2 \left(1 + P_r |f_{S_k, r}|^2\right) \leq Q_r^{\max}, \text{ for } r = 1, 2, \dots, N_r.
\end{aligned} \tag{9.28}$$

Investigating its dual problem, the authors of [22] show that the minimization problem (9.28) can be solved by iteratively solving  $N$  inner optimization problems and an outer optimization problem with  $N_r$  real dimensions, thereby proposing an efficient numerical algorithm.

Another method considered in [22] relies on the maximization of the minimum SNR margin under a total relay power constraint and under individual relay power constraints. These maximization problems can be written, respectively, as

$$\max_{\{\mathbf{w}_k\}_{k=1}^K} \min_{1 \leq k \leq K} \frac{\text{SNR}_k}{\gamma_k}, \text{ subject to } \sum_{k=1}^K P_k \|\mathbf{w}_k \odot \mathbf{b}_k\|^2 \leq Q_r^{\max} \tag{9.29}$$

and

$$\max_{\{\mathbf{w}_k\}_{k=1}^K} \min_{1 \leq k \leq K} \frac{\text{SNR}_k}{\gamma_k}, \text{ subject to } \sum_{k=1}^K |w_{S_k, r}|^2 \left(1 + P_r |f_{S_k, r}|^2\right) \leq Q_r^{\max}, \text{ for } r = 1, 2, \dots, N_r. \tag{9.30}$$

For a given SNR margin, both problems are transformed into convex feasibility problems, and thus, they can be solved by a combination of bisection search and second-order cone programming. In [22], two other approaches are reported to solve the optimization problem (9.29). First, based on [Result 9.1](#), this problem can be simplified into a convex optimization over the power allocation among different source-destination pairs. Second, the problem with its simplification based on [Result 9.1](#) is shown to be the inverse problem of the power minimization problem explained previously, thus can be solved via the solution of the power minimization problem. For the optimization problem (9.30), similar to the optimization problem (9.24), an iterative optimization scheme is proposed by studying its dual problem.

In [23], the authors consider the problem of network sum-rate maximization under a total relay power constraint. This problem can be written as

$$\max_{\{\mathbf{w}_k\}_{k=1}^K} \sum_{k=1}^K \log(1 + \text{SNR}_k), \text{ subject to } \sum_{k=1}^K P_k \|\mathbf{w}_k \odot \mathbf{b}_k\|^2 \leq Q_T^{\max}. \tag{9.31}$$

Based on [Result 9.1](#), the sum-rate maximization problem (9.31), which is over a space of  $KN_r$  complex dimensions, can be transformed to another problem over a space of only  $K$  real dimensions. The new problem can then be shown to be convex and can be solved exactly using the water-filling solution. Furthermore, the SNR and sum-rate of the network with the proposed solution are analyzed when the number of

relays  $N_r$  is asymptotically large. The derived analytical results provide the performance scaling of the network for high SNR regime.

In [24], for the same network considered in [23], the authors study the minimization of the maximum relay power subject to received SNR constraint for all source-destination pairs. With Eqs. (9.23) and (9.26), this minimization problem is cast as

$$\min_{\{\mathbf{w}_k\}_{k=1}^K} \max_{1 \leq r \leq N_r} \sum_{k=1}^K |w_{S_k,r}|^2 \left( 1 + P_r |f_{S_k,r}|^2 \right), \text{ subject to } \frac{P_k |\mathbf{w}_k^H (\mathbf{f}_k \odot \mathbf{g}_k)|^2}{1 + \|\mathbf{w}_k \odot \mathbf{g}_k\|^2} \geq \gamma_k, \text{ for } k = 1, 2, \dots, K. \quad (9.32)$$

The authors of [24] prove that for the minimization problem (9.32), the duality gap is zero, and then by reformatting the dual problem into an SDP problem, they propose a numerical algorithm to solve the problem whose worst complexity is shown to be polynomial in the number of relays and the number of source-destination pairs. Also considered in [24] is the beamforming design which maximizes the minimum received SNR under individual relay power constraints. This maximization can be written as

$$\begin{aligned} & \max_{\{\mathbf{w}_k\}_{k=1}^K, \gamma} \gamma \\ & \text{subject to } \text{SNR}_k \geq \gamma \text{ and } \sum_{k=1}^K |w_{S_k,r}|^2 \left( 1 + P_r |f_{S_k,r}|^2 \right) \leq Q_r^{\max}, \text{ for } r = 1, 2, \dots, N_r. \end{aligned} \quad (9.33)$$

An iterative algorithm to solve the maximization problem (9.33) is proposed via solving the per-relay power minimization with bisection search on the maximum per-relay power target.

### 9.3.1.8 With user interference and perfect CSI

As the number of users and demands for data rates increase drastically, multi-user communication schemes that allow several users to send information using the same time and frequency resources become more and more important. As a result, research activities on relay beamforming designs are growing for systems where both the relay channels and the user channels are nonorthogonal. Only one relay beamforming vector,  $\mathbf{w}$  is used to serve all source-destination pairs simultaneously.

Considering such a relay network, the investigations in [25–28] assume perfect CSI. Based on the two-step relaying protocol, the received signal-to-interference-plus-noise-ratio (SINR) at Destination  $k$  can be written as

$$\text{SINR}_k = \frac{P_k |\mathbf{w}^H (\mathbf{f}_k \odot \mathbf{g}_k)|^2}{1 + \|\mathbf{w} \odot \mathbf{g}_k\|^2 + \sum_{m=1, m \neq k}^K P_m |\mathbf{w}^H (\mathbf{f}_m \odot \mathbf{g}_k)|^2}.$$

The transmit power of Relay  $r$  can be expressed as

$$Q_r = |w_r|^2 \left( 1 + \sum_{k=1}^K P_k |f_{S_k,r}|^2 \right).$$

The total power consumed by all relays is thus given by

$$Q_T = \sum_{r=1}^{N_r} Q_r = \|\mathbf{w}\|^2 + \sum_{k=1}^K P_k \|\mathbf{w} \odot \mathbf{f}_k\|^2.$$

In [25, 26], the relay beamforming vector is designed to maximize the smallest of the information throughputs at all destinations under individual relay power constraints. This maximization problem is formulated as

$$\begin{aligned} & \max_{\mathbf{w}} \min_{1 \leq k \leq K} \log(1 + \text{SINR}_k), \\ \text{subject to } & Q_r = |w_r|^2 \left( 1 + \sum_{k=1}^K P_k |f_{S_k, r}|^2 \right) \leq Q_r^{\max}, \text{ for } r = 1, 2, \dots, N_r. \end{aligned} \quad (9.34)$$

By showing that the objective function in the optimization problem (9.34) is a difference-of-convex function, a low-complexity iterative algorithm is developed to find a relay beamforming solution. In [26], the relay beamforming vector is designed via the minimization of the total relay power under both the individual relay power constraints and the SINR constraints. This minimization amounts to solving the following optimization problem:

$$\min_{\mathbf{w}} \|\mathbf{w}\|^2 + \sum_{k=1}^K P_k \|\mathbf{w} \odot \mathbf{f}_k\|^2, \quad (9.35)$$

subject to  $\text{SINR}_k \geq \gamma_k$ , for  $k = 1, 2, \dots, K$ , and  $Q_r \leq Q_r^{\max}$ , for  $r = 1, 2, \dots, N_r$ .

To solve the optimization problem (9.35), one can use a non-smooth optimization algorithm, which takes advantage of the difference-of-convex-function structure of the objective function in the optimization problem (9.35). Networks with orthogonal user transmissions are also studied in [26].

In [27], joint relay beamforming vector design and relay assignment is studied by maximizing the minimum SINR subject to individual and total relay power constraints. Both cases of orthogonal and nonorthogonal transmissions for different pairs are considered. The problems are solved efficiently using the framework of difference-of-convex-function optimization. The work in [28] conducts convergence analysis for various adaptive relay beamforming schemes which can be reformulated within the random search framework. Two sufficient conditions were derived for guaranteed convergence.

### 9.3.1.9 With user interference and partial CSI

The authors of [29] study the relay beamforming design and performance analysis for multi-user multi-relay networks with user interference under quantized CSI. The CSI quantizer is designed to maximize the probability that all destinations correctly decode their desired symbol, thereby minimizing the error rate of the vector of all user signals. More importantly, to accommodate the complicated behavior of the multi-user network, a generalized diversity measure, that has two orders, is introduced in [29]. The first-order diversity coincides with the traditional diversity measure, while the second-order diversity is concerned with the transmit power

dependent logarithmic terms that appear in the error rate expression. Rigorous diversity analysis for relay beamforming with both AF and decode-and-forward (DF) schemes are also conducted.

In [30], relay beamforming design is investigated for multi-user relay networks with second-order statistical CSI only. The total relay power is minimized subject to SINR constraints for all source-destination pairs. The problem is solved using SDP. Another work that uses second-order statistical CSI only is [31]. The authors of [31] consider a multi-cluster network, where in each cluster, multi-user peer-to-peer communications are established with the help of multiple single-antenna relays. The communications of different clusters interfere with each other. The relay beamforming coefficients are optimized to minimize the total relay power subject to SINR constraints. A computationally efficient solution is obtained first by relaxing the problem to an SDP problem. Based on this relaxed problem, distributed algorithms are then proposed, where each cluster only needs to exchange information with neighboring clusters.

### 9.3.1.10 Robust designs against CSI errors

In practice, the CSI is always subject to error due to many factors such as imperfect synchronization and noises. Robust beamforming designs that aim to accommodate CSI errors are considered in [32–36]. In [33], robust relay beamforming is designed by minimizing the total relay power subject to SINR requirements on all source-destination pairs for all CSI within a fixed perturbation bound. This minimization problem is relaxed into a semidefinite problem and solved using interior point methods. In [32], imperfect CSI is modeled by adding Gaussian perturbation to the nominal (or presumed) channel coefficients. The relay transmit power is then minimized subject to outage probability constraints. To solve the problem, the asymptotic case with a large number of relays is considered, thus central limit theorem (CLT) is applied. Along with semidefinite relaxation, the problem is transformed to a convex one, based on which three suboptimal methods are proposed. For the same network model and problem, in [34, 35], approximations are made to reformulate the nonconvex probabilistic constraint, based on which the problem is solved through semidefinite relaxation and interior point method. In [36], only the relay-to-destination channels are assumed to be perturbed. For the same problem, two conservative reformulations are proposed and solved using SDP.

### 9.3.2 NETWORKS WITH FREQUENCY-SELECTIVE CHANNELS

Considered in this subsection are relay beamforming designs for networks with frequency-selective channels. Due to the frequency-selectivity of the channels, instead of using one coefficient at each relay, the FF scheme can be employed for relay processing [1]. In the FF protocol, an FIR filter is used at each relay to process the relay received signals. We denote the  $n$ th tap of the conjugate of the impulse response coefficient of the filter used at Relay  $r$  as  $w_r[n]$  (see Eq. 9.1) and represent the length of the relay filters as  $L_w$ . With slight abuse of notation used in previous sections, we define

$$\mathbf{w}_n \triangleq [w_1[n] \quad w_2[n] \quad \dots \quad w_{N_r}[n]]^T,$$

which is the relay beamforming vector of all relays on the  $n$ th tap, for  $n = 0, 1, \dots, L_w - 1$ .

### 9.3.2.1 Single-user networks

Consider the single-user network shown in Fig. 9.1, where a single source transmits information to a single destination with the help of  $N_r$  single-antenna relays. To account for the frequency-selectivity of the channels, define the following vectors

$$\begin{aligned}\check{\mathbf{f}}_n &\triangleq [\check{f}_{11}[n] \quad \check{f}_{12}[n] \quad \dots \quad \check{f}_{1N_r}[n]]^T, \text{ for } n=0,1,\dots,L_f-1, \\ \check{\mathbf{g}}_n &\triangleq [\check{g}_{12}[n] \quad \check{g}_{22}[n] \quad \dots \quad \check{g}_{N_r2}[n]]^T, \text{ for } n=0,1,\dots,L_g-1,\end{aligned}$$

which are the  $n$ th taps of the impulse responses of the source-to-relay and the relay-to-destination channels, while  $L_f$  and  $L_g$  are the lengths of the source-to-relay and the relay-to-destination CIRs, respectively.

In what follows, the two-step communication protocol and the corresponding formulations for the single-user relay network are explained. In the first step, the transmission is from the source to the relays, which can be represented as

$$\mathbf{x}(n) = \sqrt{P} \sum_{n'=0}^{L_f-1} \mathbf{f}_{n'} s(n-n') + \mathbf{v}(n),$$

where  $P$  is the source transmit power,  $\mathbf{x}(n)$  is the  $N_r \times 1$  vector of the signals received at the relays in the  $n$ th channel use,  $s(n)$  is the signal transmitted by the source in the  $n$ th channel use,  $\mathbf{v}(n)$  is the noise vector at the relays in the  $n$ th channel use. The information symbols are assumed to be independent with zero-mean and unit-variance. Recall that with FF [1], an FIR filter is used at each relay and  $\mathbf{w}_l$  is the relay beamforming vector on the  $l$ th tap. With slight abuse of notation used in previous sections, define

$$\mathbf{w} \triangleq [\mathbf{w}_0^T \quad \mathbf{w}_1^T \quad \dots \quad \mathbf{w}_{L_w-1}^T]^T,$$

which contains the relay coefficients at all relays across all taps. The signal vector sent from the relays to the destination at the second step on the  $n$ th channel use can be expressed as

$$\mathbf{t}(n) = \sum_{n'=0}^{L_w-1} \mathbf{w}_{n'}^* \odot \mathbf{x}(n-n').$$

The following notation is introduced to help the formulation. Let

$$\mathbf{B}_m \triangleq \begin{bmatrix} \mathbf{0}_{m-1,1} & \mathbf{I}_{m-1,m-1} \\ 1 & \mathbf{0}_{1,m-1} \end{bmatrix},$$

where  $\mathbf{0}_{ij}$  is the  $i \times j$  matrix of all zeros.  $\mathbf{B}_m$  is an  $m \times m$  permutation matrix. By right multiplying any matrix with  $\mathbf{B}_m$ , the columns of that matrix are circularly shifted to the left by 1. Define

$$\begin{aligned}\mathbf{F}_0 &\triangleq \begin{bmatrix} \check{\mathbf{f}}_0 & \dots & \check{\mathbf{f}}_{L_f-1} & \mathbf{0}_{N_r, L_w-1} \end{bmatrix}, \quad \check{\mathbf{F}} \triangleq \begin{bmatrix} \mathbf{F}_0^T & (\mathbf{F}_0 \mathbf{B}_{L_f+L_w-1})^T & \dots & (\mathbf{F}_0 \mathbf{B}_{L_f+L_w-1}^{L_w-1})^T \end{bmatrix}^T, \\ \check{\mathbf{F}}_0 &\triangleq \begin{bmatrix} \check{\mathbf{F}} & \mathbf{0}_{N_r, L_w, L_g-1} \end{bmatrix}, \quad \check{\mathbf{F}} \triangleq \begin{bmatrix} \check{\mathbf{F}}_0^T & (\check{\mathbf{F}}_0 \mathbf{B}_{L_f+L_w+L_g-2})^T & \dots & (\check{\mathbf{F}}_0 \mathbf{B}_{L_f+L_w+L_g-2}^{L_g-1})^T \end{bmatrix}^T, \\ \check{\mathbf{I}}_0 &\triangleq \begin{bmatrix} \mathbf{I}_{N_r, L_w} & \mathbf{0}_{N_r, L_w, (L_g-1)} \end{bmatrix}, \quad \check{\mathbf{I}} \triangleq \begin{bmatrix} \mathbf{I}_0^T & (\check{\mathbf{I}}_0 \mathbf{B}_{N_r, (L_f+L_w+L_g-2)})^T & \dots & (\check{\mathbf{I}}_0 \mathbf{B}_{N_r, (L_f+L_w+L_g-2)}^{L_g-1})^T \end{bmatrix}^T.\end{aligned}$$

After some manipulations, the received signal at the destination can be expressed as

$$y(n) = \sqrt{P} \mathbf{w}^H \mathfrak{G} \check{\mathbf{F}} \check{\mathbf{s}}(n) + \mathbf{w}^H \mathfrak{G} \check{\mathbf{I}} \check{\mathbf{v}}(n) + u_d(n), \quad (9.36)$$

where

$$\begin{aligned}\check{\mathbf{s}}(n) &\triangleq [s(n) \quad s(n-1) \quad \dots \quad s(n-L_f-L_w-L_g+3)]^T, \\ \check{\mathbf{v}}(n) &\triangleq [\mathbf{v}^T(n) \quad \mathbf{v}^T(n-1) \quad \dots \quad \mathbf{v}^T(n-L_w-L_g+2)]^T, \\ \mathfrak{G} &\triangleq [\mathbf{I}_{L_w} \otimes \text{diag}\{\check{\mathbf{g}}_0\} \quad \dots \quad \mathbf{I}_{L_w} \otimes \text{diag}\{\check{\mathbf{g}}_{L_g-1}\}],\end{aligned}$$

and  $u_d(n)$  is the destination noise at the  $n$ th channel use. Note that the first term in Eq. (9.36) contains both the desired signal for the  $n$ th channel use  $s(n)$  (the first element of  $\check{\mathbf{s}}(n)$ ) and the ISI term caused by other information symbols. Thus the SINR (which is now the signal-to-ISI-plus-noise ratio), instead of the SNR, determines the performance of the network. By decomposing the matrix  $\check{\mathbf{F}}$  as  $\check{\mathbf{F}} \triangleq [\bar{\mathbf{f}} \quad \bar{\mathbf{F}}]$ , where  $\bar{\mathbf{f}}$  is the first column of  $\check{\mathbf{F}}$ , after some calculations, the received SINR can be expressed as

$$\text{SINR} = \frac{\mathbf{w}^H \mathbf{Q}_s \mathbf{w}}{\mathbf{w}^H \mathbf{Q}_i \mathbf{w} + \mathbf{w}^H \mathbf{Q}_n \mathbf{w} + 1},$$

where

$$\begin{aligned}\mathbf{Q}_s &\triangleq P [\mathbf{I}_{N_r} \quad \mathbf{0}_{N_r, (L_w-1)N_r}]^H (\check{\mathbf{f}}_0 \odot \check{\mathbf{g}}_0) (\check{\mathbf{f}}_0 \odot \check{\mathbf{g}}_0)^H [\mathbf{I}_{N_r} \quad \mathbf{0}_{N_r, (L_w-1)N_r}], \\ \mathbf{Q}_i &\triangleq P \mathfrak{G} \bar{\mathbf{F}} \bar{\mathbf{F}}^H \mathfrak{G}^H, \\ \mathbf{Q}_n &\triangleq \mathfrak{G} \tilde{\mathbf{I}} \tilde{\mathbf{I}}^H \mathfrak{G}^H.\end{aligned}$$

The transmit power of the  $r$ th relay can be calculated to be

$$Q_r = \mathbf{w}^H \mathfrak{D}_r \mathbf{w},$$

where

$$\mathfrak{D}_r \triangleq P_s (\mathbf{I}_{L_w} \otimes \mathfrak{E}_r) (\check{\mathbf{F}} \check{\mathbf{F}}^H + \mathbf{I}_{N_r L_w}) (\mathbf{I}_{L_w} \otimes \mathfrak{E}_r)^H$$

with  $\mathfrak{E}_r$  being the matrix with all zeros entries except for the  $(r, r)$ th entry, which is 1. The total relay power can then be represented as

$$Q_T = \sum_{r=1}^{N_r} Q_r = \mathbf{w}^H \mathfrak{D} \mathbf{w},$$

where

$$\mathfrak{D} = \sum_{r=1}^{N_r} \mathfrak{D}_r = P_s \sum_{r=1}^{N_r} (\mathbf{I}_{L_w} \otimes \mathbf{E}_r) \mathfrak{F} \mathfrak{F}^H (\mathbf{I}_{L_w} \otimes \mathbf{E}_r)^H + \mathbf{I}_{RL_w}.$$

Three relay beamforming design problems are investigated in [1]. The first problem is the relay power minimization under the SINR constraint  $\gamma$ . This problem is formulated as

$$\min_{\mathbf{w}} Q_T, \text{ subject to } \text{SINR} \geq \gamma,$$

or, equivalently, as

$$\min_{\mathbf{w}} \mathbf{w}^H \mathfrak{D} \mathbf{w}, \text{ subject to } \frac{\mathbf{w}^H \mathbf{Q}_s \mathbf{w}}{\mathbf{w}^H \mathbf{Q}_i \mathbf{w} + \mathbf{w}^H \mathbf{Q}_n \mathbf{w} + 1} \geq \gamma.$$

Based on Lagrange multiplier method, a closed-form solution to this problem can be found, as summarized below.

### RESULT 9.8

For single-user multi-relay networks with frequency-selective channels, the optimal FF relay beamforming design which minimizes the total relay transmit power subject to a feasible received SINR constraint  $\gamma$  is

$$\mathbf{w}^o = \left( \frac{\gamma}{[\mathcal{P}(\mathbf{C})]^H \mathcal{C} \mathcal{P}(\mathbf{C})} \right)^{1/2} \mathfrak{D}^{-1/2} \mathcal{P}(\mathbf{C}),$$

where  $\mathbf{w}$  is the vector containing the impulse response coefficients of all relay filters and

$$\mathbf{C} \triangleq \mathfrak{D}^{-1/2} (\mathbf{Q}_s - \gamma \mathbf{Q}_i - \gamma \mathbf{Q}_n) \mathfrak{D}^{-1/2}.$$

With this optimal design, the minimum total relay transmit power is

$$Q_T^o = \gamma / \lambda_{max}(\mathbf{C}).$$

The second design is the SINR maximization under the total relay power constraint, formulated as

$$\max_{\mathbf{w}} \text{SINR}, \text{ subject to } Q_T \leq Q_T^{\max},$$

which can be specified as

$$\max_{\mathbf{w}} \frac{\mathbf{w}^H \mathbf{Q}_s \mathbf{w}}{\mathbf{w}^H \mathbf{Q}_i \mathbf{w} + \mathbf{w}^H \mathbf{Q}_n \mathbf{w} + 1}, \text{ subject to } \mathbf{w}^H \mathfrak{D} \mathbf{w} \leq Q_T^{\max}.$$

A closed-form solution to this problem can be found as summarized in [Result 9.9](#).

**RESULT 9.9**

For single-user multi-relay networks with frequency-selective channels, the optimal FF relay beamforming design that maximizes the received SINR for a given total relay transmit power constraint  $\mathcal{Q}_T^{\max}$ , is given by

$$\mathbf{w}^o = \sqrt{\mathcal{Q}_T^{\max}} \mathfrak{D}^{-1/2} \mathcal{P} \left( \left[ \frac{1}{\mathcal{Q}_T^{\max}} \mathbf{I} + \mathfrak{D}^{-1/2} (\mathbf{Q}_i + \mathbf{Q}_n) \mathfrak{D}^{-1/2} \right]^{-1} \mathfrak{D}^{-1/2} \mathbf{Q}_s \mathfrak{D}^{-1/2} \right).$$

With this optimal design, the achieved maximum received SINR is

$$\text{SINR}^o = \lambda_{\max} \left( \left[ \frac{1}{\mathcal{Q}_T^{\max}} \mathbf{I} + \mathfrak{D}^{-1/2} (\mathbf{Q}_i + \mathbf{Q}_n) \mathfrak{D}^{-1/2} \right]^{-1} \mathfrak{D}^{-1/2} \mathbf{Q}_s \mathfrak{D}^{-1/2} \right).$$

The third design is the SINR maximization under individual per-relay power constraints. This design problem can be formulated as

$$\max_{\mathbf{w}} \text{SINR}, \text{ subject to } Q_r \leq \mathcal{Q}_r^{\max}, \text{ for } r = 1, 2, \dots, N_r,$$

or, equivalently, as

$$\max_{\mathbf{w}} \frac{\mathbf{w}^H \mathbf{Q}_s \mathbf{w}}{\mathbf{w}^H \mathbf{Q}_i \mathbf{w} + \mathbf{w}^H \mathbf{Q}_n \mathbf{w} + 1}, \text{ subject to } \mathbf{w}^H \mathfrak{D}_r \mathbf{w} \leq \mathcal{Q}_r^{\max}, \text{ for } r = 1, 2, \dots, N_r.$$

Due to the challenge in dealing with individual power constraints, closed-form solution to this problem is not available. Instead, the problem can be transformed into a quasiconvex optimization, and thus, can be solved using a numerical algorithm which combines a bisection search and a second-order cone feasibility problem.

In [3], the relay filter design for the FF scheme is studied with an additional linear or decision feedback equalization at the destination. In this work, infinite impulse response (IIR) filters are allowed at the relays. The relay filter design which maximizes the SINR at the output of the equalizer subject to the total relay power constraint is studied. Different from the time-domain approach in [1], the formulation and the solution in [3] are in frequency-domain. For IIR filters, the filter frequency response expression (that is valid for linear equalization, decision feedback equalization), and an idealized matched filter receiver are obtained. The expression depends on a power allocation factor, which can be found via numerical optimization. For FIR filters at the relays, a gradient-based algorithm is proposed. In [37], the relay beamforming problem is also investigated via frequency domain analysis. The authors aim to find the SINR maximizing relay filters under a total relay transmit power and another constraint where the equivalent channel from the source to the destination is distortionless. A closed-form solution to the relay beamforming filters is found. The adaptive selection of the decision delay that achieves the highest output SINR is also derived.

### 9.3.2.2 Multi-user networks

In [38], the FF relay beamforming is extended from single-user to multi-user relay networks. In [38], the relay beamforming filters are designed to maximize the worst received SINR among all source-destination pairs subject to total and/or individual relay power constraints. A semidefinite relaxation can be used to turn the problem into a combination of SDP methods and a bisection search. A more recent development on relay beamforming for multi-user multi-relay networks with frequency-selective channels can be found in [39], where individual adaptive decoding delays at the destinations are further considered to effectively mitigate user-interference. The FIR filter coefficients at the relay and the decoding delays at the destinations are jointly optimized to minimize the total relay power subject to SINR constraints at all destinations. In [40], the source power allocation and relay beamforming are jointly designed in multi-user peer-to-peer multi-relay networks that use single-carrier frequency division multiple access for frequency-selective channels. Through difference-of-convex-function program, numerical algorithms are provided for two optimization problems: maximizing worst SINR subject to various power constraints and minimizing total transmit power subject to SINR constraints.

---

## 9.4 TWO-WAY NETWORK BEAMFORMING

To enable bidirectional relay-assisted communication between two transceivers, a straightforward approach is to use two consecutive one-way relaying schemes, each of which allows the flow of information symbols from one transceiver to the other one. As each one-way relaying scheme takes two time-slots to transport one information symbol from one transceiver to the other one, this scheme requires four time-slots to materialize a two-way symbol exchange between the two transceivers. As a more bandwidth efficient alternative, a three-time-slot relaying scheme, namely the time division broad cast (TDBC) technique, can be used to enable such an information exchange. In the TDBC approach, the two transceivers transmit, to the relays, their information symbols in the first two temporally orthogonal time-slots. In the third time-slot, each relay linearly combines the signals it receives in the first two time-slots and retransmits the so-obtained signal toward the transceivers. The complex coefficients used to combine the relays' received signals constitute the design parameters which have to be obtained through optimal design.

The multiple access broadcast channel (MABC) scheme is yet another relaying scheme used to facilitate a two-way communication between two transceivers. In this two-time-slot scheme, the transceivers simultaneously transmit, in the first time-slot, their information symbols to the relays. Each relay receives a faded noisy mixture of the two signals transmitted by the two transceivers. Using a complex beamforming weight, each relay then judiciously adjusts the phase and the amplitude of its received signal and retransmits the so-obtained signal. Each transceiver first cancels the self-interference from its received signal and then extracts its symbol of interest from the residual signal. It has been shown that in the absence of strong direct link between the

two transceivers, the MABC scheme outperforms the TDBC approach in terms of the total power consumed in the entire network for given SNR thresholds at the two transceivers or in terms of the maximum balanced SNR achieved under a total power budget constraint [41]. Unless otherwise stated explicitly, the rest of the main focus of this chapter will be on the two-time-slot MABC type of bidirectional relaying schemes.

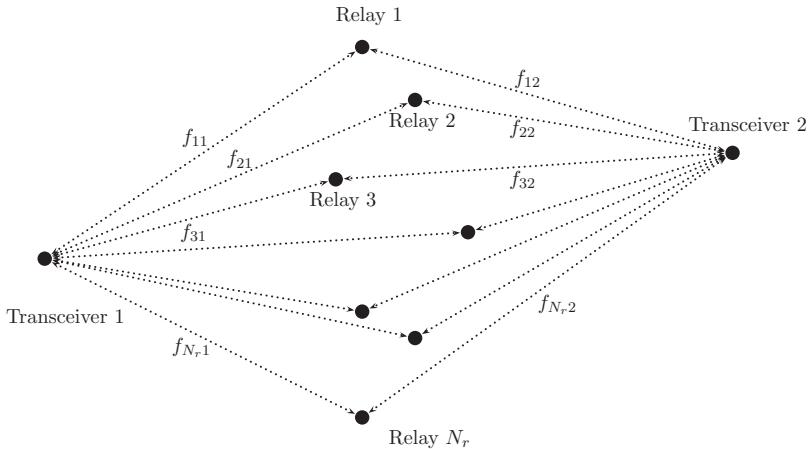
In two-way relaying, it is often assumed that the transceiver-relay channels are reciprocal, that is  $\check{f}_{qr}[\cdot] = \check{g}_{rq}[\cdot]$ , for  $q = 1, 2, \dots, 2N_p$  and  $r = 1, 2, \dots, N_r$ . The channel reciprocity holds for sufficiently slowly varying wireless transceiver-relay channels, when the gain of each node's transmission filter is the same as the gain of that node's receive filter.<sup>2</sup> In this section, different designs of two-way network beamforming schemes under different assumptions are presented. First, the focus will be on networks with frequency-flat transceiver-relay links, implying that the CIRs of the reciprocal link between Transceiver  $q$  and the  $r$ th relay link can be written as  $\check{f}_{qr}[n] = f_{rq}\delta[n - \check{n}_{rq}]$ , where  $\delta[n]$  is 1 for  $n = 0$  and zero otherwise,  $\check{n}_{rq}$  represents the time delay (in terms of the number of symbols) of the link between the  $r$ th relay and Transceiver  $q$  and  $f_{rq}$  is the corresponding flat fading channel coefficient, for  $q = 1, 2, \dots, 2N_p$  and  $r = 1, 2, \dots, N_r$ . For networks with frequency-flat transceiver-relay links both cases of synchronous and asynchronous schemes are presented in the next two subsections. Networks with frequency-selective transceiver-relay channels are discussed in the third subsection.

#### 9.4.1 SYNCHRONOUS NETWORKS

Shown in Fig. 9.3 is a wireless relay network which consists of two single-antenna transceivers ( $N_p = 1$ ) and  $N_r$  single-antenna relay nodes. The network is assumed to be synchronous, i.e.,  $\check{n}_{rq} = n_q$ , for  $q = 1, 2$ , and  $r = 1, 2, \dots, N_r$ , where  $n_q$  is the delay of signal propagation from/to each transceiver to/from different relays. In other words, the delays of signal propagation from/to each transceiver to/from different relays are identical (or within one symbol period). This assumption implies that the signals transmitted by each transceiver arrive at different relays with the same delay and the signals transmitted by different relays also arrive at each transceiver with the same delay. In this case, the end-to-end CIR in Eq. (9.1) can be written as

---

<sup>2</sup>Note that the end-to-end channel reciprocity depends not only on the wireless channel but also on the receiver and transmitter filters. The link from a transceiver to a relay involves the transmit filter of the transceiver and the receive filter of the relay, while the link from a relay to a transceiver involves the transmit filter of the relay and the receive filter of the transceiver. In order for a transceiver-relay link to satisfy the channel reciprocity condition not only does the wireless channel between the transceiver and the relay have to be reciprocal, but also the combined effect of the transmit filter of the transceiver and the receive filter of the relay should be the same as the combined effect of the transmit filter of the relay and received filter of the transceiver. One way to guarantee this condition is to ensure that the gain of each node's transmit filter is the same as the gain of that node's receive filter.

**FIG. 9.3**

A two-way relay network with two transceivers [42].

$$h[n] \triangleq h_{12}[n] = h_{21}[n] = \sum_{r=1}^{N_r} f_{r1} f_{r2} w_r^* \delta[n - (\check{n}_1 + \check{n}_2)] = \mathbf{w}^H (\mathbf{f}_1 \odot \mathbf{f}_2) \delta[n - (\check{n}_1 + \check{n}_2)], \quad (9.37)$$

where  $\mathbf{f}_1 \triangleq [f_{11} \ f_{21} \ \dots \ f_{N_r1}]^T$  and  $\mathbf{f}_2 \triangleq [f_{12} \ f_{22} \ \dots \ f_{N_r2}]^T$  are the vectors of the coefficients of the links between the relays and the transceivers. As can be seen from Eq. (9.37), the end-to-end channel is also frequency-flat and reciprocal. It is herein assumed that due to the poor quality of the channel between the two transceivers, there is no direct link between them. As a result, the two transceivers communicate with each other through the relay nodes. Each relay has a single antenna for both transmission and reception.

In the first time-slot of the two-time-slot MABC method, Transceivers 1 and 2 simultaneously transmit their information symbols \$s\_1\$ and \$s\_2\$, respectively, with power \$P\_1\$ and \$P\_2\$ to the relays. Under the aforementioned assumptions, the \$N\_r \times 1\$ vector \$\mathbf{x}\$ of the relay received signals is given by

$$\mathbf{x} = \sqrt{P_1} \mathbf{f}_1 s_1 + \sqrt{P_2} \mathbf{f}_2 s_2 + \boldsymbol{\nu}, \quad (9.38)$$

where \$\boldsymbol{\nu}\$ is the \$N\_r \times 1\$ complex vector of the relay noises. Each transceiver is assumed to have the perfect knowledge of the channel vectors \$\mathbf{f}\_1\$ and \$\mathbf{f}\_2\$. In the second step, the \$r\$th relay multiplies its received signal by a complex weight \$w\_r^\*\$ and transmits the so-obtained signal. The \$N\_r \times 1\$ complex vector \$\mathbf{t}\$ of the relay transmitted signals can then be expressed as

$$\mathbf{t} = \mathbf{w}^* \odot \mathbf{x}, \quad (9.39)$$

where  $\odot$  is the Schur-Hadamard (element-wise) vector product and the relay beamforming vector is defined earlier as  $\mathbf{w} \triangleq [w_1 \ w_2 \ \dots \ w_{N_r}]^T$ . Based on Eqs. (9.38) and (9.39), the transceivers' received signals  $y_1$  and  $y_2$  are then written as

$$y_1 = \mathbf{f}_1^T (\mathbf{w}^* \odot \mathbf{x}) + v_1 = \mathbf{f}_1^T (\mathbf{w}^* \odot (\sqrt{P_1} \mathbf{f}_1 s_1 + \sqrt{P_2} \mathbf{f}_2 s_2 + \mathbf{v})) + v_1, \quad (9.40)$$

$$\begin{aligned} &= \sqrt{P_1} \mathbf{w}^H (\mathbf{f}_1 \odot \mathbf{f}_1) s_1 + \sqrt{P_2} \mathbf{w}^H (\mathbf{f}_1 \odot \mathbf{f}_2) s_2 + \mathbf{w}^H \mathbf{f}_1 \mathbf{v} + v_1, \\ y_2 &= \mathbf{f}_2^T (\mathbf{w}^* \odot \mathbf{x}) + v_2 = \mathbf{f}_2^T (\mathbf{w}^* \odot (\sqrt{P_1} \mathbf{f}_1 s_1 + \sqrt{P_2} \mathbf{f}_2 s_2 + \mathbf{v})) + v_2 \\ &= \sqrt{P_1} \mathbf{w}^H (\mathbf{f}_2 \odot \mathbf{f}_1) s_1 + \sqrt{P_2} \mathbf{w}^H (\mathbf{f}_2 \odot \mathbf{f}_2) s_2 + \mathbf{w}^H \mathbf{f}_2 \mathbf{v} + v_2, \end{aligned} \quad (9.41)$$

where  $v_q$  is the received noise at Transceiver  $q$ , for  $q = 1, 2$ . In this scheme, both transceivers calculate the optimal values of the weight vector and the optimum transceiver transmit powers. Note that the first term in Eq. (9.40), referred to as self-interference, depends on the signal  $s_1$  transmitted by Transceiver 1 during the first time-slot, on the vector  $\sqrt{P_1} \mathbf{f}_1 \odot \mathbf{f}_1$ , which is known to Transceiver 1, and on the weight vector  $\mathbf{w}$ , which is to be calculated at this transceiver. Hence, the self-interference term in Eq. (9.40) is known at Transceiver 1, and thus, this term can be subtracted from  $y_1$ . As a result the residual signal can be processed at Transceiver 1 to extract the information symbol  $s_2$ . Similarly, the second term in Eq. (9.41) is the self-interference term at Transceiver 2 and can be subtracted from  $y_2$ . The residual signal can be processed at Transceiver 2 with the aim to extract the information symbol  $s_1$ . More specifically, the residual signals  $\tilde{y}_1$  and  $\tilde{y}_2$ , defined as

$$\tilde{y}_1 \triangleq y_1 - \underbrace{\sqrt{P_1} \mathbf{w}^H (\mathbf{f}_1 \odot \mathbf{f}_1) s_1}_{\text{desired signal at Transceiver 1}} - \underbrace{\mathbf{w}^H (\mathbf{f}_1 \odot \mathbf{v}) + v_1}_{\text{noise}}, \quad (9.42)$$

$$\tilde{y}_2 \triangleq y_2 - \underbrace{\sqrt{P_2} \mathbf{w}^H (\mathbf{f}_2 \odot \mathbf{f}_2) s_2}_{\text{desired signal at Transceiver 2}} - \underbrace{\mathbf{w}^H (\mathbf{f}_2 \odot \mathbf{v}) + v_2}_{\text{noise}} \quad (9.43)$$

are processed at their corresponding transceivers to detect the respective desired information symbols. The noise process is assumed to be zero-mean and spatially white with variance  $\sigma^2$ . That is,  $E\{|v_1|^2\} = E\{|v_2|^2\} = \sigma^2$  and  $E\{\mathbf{v}\mathbf{v}^H\} = \sigma^2 \mathbf{I}$ . Defining  $P_q$  as the transmit power of Transceiver  $q$  implies that  $E\{|s_q|^2\} = 1$  holds true, for  $q = 1, 2$ .

To obtain the design parameters, namely the transceivers' transmit powers  $P_1$  and  $P_2$  as well as the relay weight vector  $\mathbf{w}$ , numerous approaches have been considered in the literature. Three popular approaches are (1) a total power minimization method subject to SNR (or rate) constraints, (2) a max-min fair SNR approach, where the smaller of the two transceivers' SNRs is maximized subject to a total power constraint, and (3) a sum-rate maximization technique subject to a total power constraint. The next subsections present the results obtained based on these three approaches. To do so, the expression for the total transmit power consumed in the entire network and

those for the transceivers' SNRs are next presented in terms of the design parameters. The transmit power  $P_T$  consumed in the entire network can be written as

$$P_T = P_1 + P_2 + E\{\mathbf{t}^H \mathbf{t}\} = P_1(1 + \mathbf{w}^H \mathbf{D}_1 \mathbf{w}) + P_2(1 + \mathbf{w}^H \mathbf{D}_2 \mathbf{w}) + \sigma^2 \mathbf{w}^H \mathbf{w}, \quad (9.44)$$

where  $\mathbf{D}_1 \triangleq \text{diag}(\mathbf{f}_1 \odot \mathbf{f}_1^*)$  and  $\mathbf{D}_2 \triangleq \text{diag}(\mathbf{f}_2 \odot \mathbf{f}_2^*)$  are two  $N_r \times N_r$  diagonal matrices, and Eq. (9.39) has been used along with the assumptions  $E\{\mathbf{v}\mathbf{v}^H\} = \sigma^2 \mathbf{I}_{N_r}$  and  $E\{|s_q|^2\} = 1$ , for  $q = 1, 2$ , to obtain  $E\{\mathbf{t}^H \mathbf{t}\} = \mathbf{w}^H (P_1 \mathbf{D}_1 + P_2 \mathbf{D}_2 + \sigma^2 \mathbf{I}_{N_r}) \mathbf{w}$ . Using Eqs. (9.42) and (9.43), the transceivers' received SNRs can be written as

$$\text{SNR}_1(P_2, \mathbf{w}) = \frac{P_2 \mathbf{w}^H \mathbf{a} \mathbf{a}^H \mathbf{w}}{\sigma^2 (1 + \mathbf{w}^H \mathbf{D}_1 \mathbf{w})}, \quad \text{SNR}_2(P_1, \mathbf{w}) = \frac{P_1 \mathbf{w}^H \mathbf{a} \mathbf{a}^H \mathbf{w}}{\sigma^2 (1 + \mathbf{w}^H \mathbf{D}_2 \mathbf{w})}, \quad (9.45)$$

where the  $N_r \times 1$  vector  $\mathbf{a}$  is defined as  $\mathbf{a} \triangleq \mathbf{f}_1 \odot \mathbf{f}_2 = \mathbf{f}_2 \odot \mathbf{f}_1$ , while  $\text{SNR}_q(P_{\bar{q}}, \mathbf{w})$  is the received SNR at Transceiver  $q$ , for  $q = 1, 2$ . Here,  $\bar{q} = 2$  when  $q = 1$ , and  $\bar{q} = 1$  when  $q = 2$ .

Now that the ground work has been laid, the results of the three aforementioned approaches to design jointly optimal two-way network beamformer and power allocation scheme are presented in the rest of this subsection.

#### 9.4.1.1 Total power minimization

In the total power minimization approach, the beamforming weight vector  $\mathbf{w}$  and the transceivers' transmit powers  $P_1$  and  $P_2$  are obtained such that the *total* transmit power  $P_T$  is minimized while maintaining the transceivers' received SNRs above predefined thresholds  $\gamma_1 > 0$  and  $\gamma_2 > 0$ . This approach amounts to solving the following optimization problem [42]:

$$\min_{P_1, P_2, \mathbf{w}} P_T, \quad \text{subject to } \text{SNR}_1(P_2, \mathbf{w}) \geq \gamma_1, \quad \text{SNR}_2(P_1, \mathbf{w}) \geq \gamma_2, \quad (9.46)$$

or, equivalently,

$$\begin{aligned} \min_{P_1, P_2, \mathbf{w}} \quad & P_1(1 + \mathbf{w}^H \mathbf{D}_1 \mathbf{w}) + P_2(1 + \mathbf{w}^H \mathbf{D}_2 \mathbf{w}) + \sigma^2 \mathbf{w}^H \mathbf{w} \\ \text{subject to} \quad & \frac{P_2 \mathbf{w}^H \mathbf{a} \mathbf{a}^H \mathbf{w}}{\sigma^2 (1 + \mathbf{w}^H \mathbf{D}_1 \mathbf{w})} \geq \gamma_1, \quad \frac{P_1 \mathbf{w}^H \mathbf{a} \mathbf{a}^H \mathbf{w}}{\sigma^2 (1 + \mathbf{w}^H \mathbf{D}_2 \mathbf{w})} \geq \gamma_2. \end{aligned} \quad (9.47)$$

At the optimum, one can easily observe that the inequality constraints in the optimization problem (9.47) must be satisfied with equality. Based on this observation, the transceivers' transmit powers can be written, in terms of  $\mathbf{w}$ , as

$$P_1 = \frac{\sigma^2 \gamma_2 (1 + \mathbf{w}^H \mathbf{D}_2 \mathbf{w})}{\mathbf{w}^H \mathbf{a} \mathbf{a}^H \mathbf{w}}, \quad P_2 = \frac{\sigma^2 \gamma_1 (1 + \mathbf{w}^H \mathbf{D}_1 \mathbf{w})}{\mathbf{w}^H \mathbf{a} \mathbf{a}^H \mathbf{w}}. \quad (9.48)$$

Note that in light of Eq. (9.48), for any pair  $(\gamma_1, \gamma_2)$ , the SNR constraints can always be satisfied by increasing the transceivers' transmit powers. Using Eq. (9.48) and ignoring the constant multiplicative scalar  $\sigma^2$ , one can simplify the optimization problem (9.47) as the following *unconstrained* optimization problem:

$$\min_{\mathbf{w}} (\gamma_1 + \gamma_2) \frac{(1 + \mathbf{w}^H \mathbf{D}_1 \mathbf{w})(1 + \mathbf{w}^H \mathbf{D}_2 \mathbf{w})}{\mathbf{w}^H \mathbf{a} \mathbf{a}^H \mathbf{w}} + \mathbf{w}^H \mathbf{w}. \quad (9.49)$$

The following result is proven in [42].

**RESULT 9.10**

All the local minimae of the cost function in the optimization problem (9.49) are global minimum. The corresponding minimizers of the cost function in the optimization problem (9.49) are different from each other only through a scalar phase rotation.

It follows from [Result 9.10](#) that the total power minimization problem (9.46) has a unique (up to a phase rotation) solution in terms of relay beamforming weight vector. Based on this result, a gradient descent-based algorithm is proposed in [42] to solve this optimization problem. A more computationally efficient solution to the optimization problem (9.49) is presented in [43]. To present this solution, note that the optimization problem (9.49) has an interesting property: as far as the optimal value of the relay beamforming vector  $\mathbf{w}$  is concerned, the solution remains the same for different pairs of  $\gamma_1$  and  $\gamma_2$  which add up to the same value. That is, for any realization of the transceiver-relay channels, as long as  $\gamma_1 + \gamma_2$  is constant, the optimal value of  $\mathbf{w}$  does not change regardless of the values of  $\gamma_1$  and  $\gamma_2$ . As a result, replacing  $\gamma_1$  and  $\gamma_2$  in the optimization problem (9.47) with 0 and  $\gamma_1 + \gamma_2$ , respectively, does not change the value of the optimal beamforming vector  $\mathbf{w}$ . Replacing  $\gamma_1$  with zero results in  $P_2$  being 0, thereby turning the network into a one-way relaying scheme. In other words, the power-optimal value of the beamforming relay weight vector  $\mathbf{w}$  in the two-way relay network with required SNR thresholds  $\gamma_1$  and  $\gamma_2$ , is identical to the power-optimal relay beamforming weight vector  $\mathbf{w}$  in the same network when this network is used in a one-way relaying scheme which establishes a connection from one transceiver (as the source) to the other transceiver (as the destination) with a required SNR of  $\gamma_1 + \gamma_2$  at the destination. Based on this property, the following *semi-closed-form solution* to the total power minimization problem is obtained in [43].

**RESULT 9.11**

The power-optimal beamforming weight vector for the optimization problem (9.46) is obtained as

$$\mathbf{w}^o = \kappa \sqrt{\frac{\tilde{P}_1(\gamma_1 + \gamma_2)}{\lambda(\tilde{P}_1)}} \left( (\gamma_1 + \gamma_2)\mathbf{D}_2 + \lambda(\tilde{P}_1)(\tilde{P}_1\mathbf{D}_1 + \mathbf{I}_{N_r}) \right)^{-1} \mathbf{a}, \quad (9.50)$$

where the parameter  $\kappa$  is obtained as

$$\kappa \triangleq \left( \tilde{P}_1 \mathbf{a}^H (\tilde{P}_1 \mathbf{D}_1 + \mathbf{I}) \left( (\gamma_1 + \gamma_2) \mathbf{D}_2 + \lambda(\tilde{P}_1) (\tilde{P}_1 \mathbf{D}_1 + \mathbf{I}_{N_r}) \right)^{-2} \mathbf{a} \right)^{-1/2} \quad (9.51)$$

and the parameter  $\tilde{P}_1$  is the power-optimal transmit power of the source in the one-way relaying scheme with receiver SNR threshold of  $\gamma_1 + \gamma_2$ . This parameter is obtained, using the Newton-Raphson method, as the *provably unique* solution to the following equation:

$$1 - (\gamma_1 + \gamma_2) \frac{1/\tilde{P}_1^2 - \lambda(\tilde{P}_1) \mathbf{a}^H \left( (\gamma_1 + \gamma_2) \mathbf{D}_2 + \lambda(\tilde{P}_1) (\tilde{P}_1 \mathbf{D}_1 + \mathbf{I}_{N_r}) \right)^{-2} \mathbf{D}_1 \mathbf{a}}{\lambda^2(\tilde{P}_1) \mathbf{a}^H \left( (\gamma_1 + \gamma_2) \mathbf{D}_2 + \lambda(\tilde{P}_1) (\tilde{P}_1 \mathbf{D}_1 + \mathbf{I}_{N_r}) \right)^{-2} (\tilde{P}_1 \mathbf{D}_1 + \mathbf{I}_{N_r}) \mathbf{a}} = 0,$$

which resides in the interval  $\left[\frac{(\gamma_1 + \gamma_2)}{\|\mathbf{f}_1\|^2}, +\infty\right)$ . For any given value of  $z$  in the interval  $\left[\frac{(\gamma_1 + \gamma_2)}{\|\mathbf{f}_1\|^2}, +\infty\right)$ , the parameter  $\lambda(z)$  is obtained, using the Newton-Raphson technique or the bisection method, as the *provably unique* positive root of the following nonlinear equation:

$$\sum_{r=1}^{N_r} \frac{z|f_{r1}f_{r2}|^2}{(\gamma_1 + \gamma_2)|f_{r2}|^2 + \lambda(z)(z|f_{r1}|^2 + 1)} = 1.$$

Alternatively, for any  $z \in \left[\frac{(\gamma_1 + \gamma_2)}{\|\mathbf{f}_1\|^2}, +\infty\right)$ ,  $\lambda(z)$  can be obtained as the principle eigenvalue of the matrix

$$\mathbf{A}(z) \triangleq (z\mathbf{D}_1 + \mathbf{I})^{-1/2} (z\mathbf{a}\mathbf{a}^H - (\gamma_1 + \gamma_2)\mathbf{D}_2) (z\mathbf{D}_1 + \mathbf{I}_{N_r})^{-1/2}.$$

The optimal values of the transceivers' transmit powers are then obtained from Eq. (9.48) by replacing  $\mathbf{w}$  with  $\mathbf{w}^o$  in Eq. (9.48). That is

$$P_1^o = \frac{\sigma^2 \gamma_2 (1 + \mathbf{w}^{o,H} \mathbf{D}_2 \mathbf{w}^o)}{\mathbf{w}^{o,H} \mathbf{a} \mathbf{a}^H \mathbf{w}^o}, \quad P_2^o = \frac{\sigma^2 \gamma_1 (1 + \mathbf{w}^{o,H} \mathbf{D}_1 \mathbf{w}^o)}{\mathbf{w}^{o,H} \mathbf{a} \mathbf{a}^H \mathbf{w}^o}.$$

Note that the parameter  $\tilde{P}_1$  is the power of Transceiver 1 normalized to the noise power in a one-way relaying scheme which establishes a connection from Transceiver 1 (as the source) to Transceiver 2 (as the destination) with a required SNR of  $\gamma_1 + \gamma_2$  at the destination. The following result is also proven in [42].

### RESULT 9.12

For any realizations of the channels, when  $\gamma_1 = \gamma_2$ , the half of the total minimum power is consumed in the relays collectively, and the remaining half is shared between the two transceivers.

[Result 9.12](#) suggests that the maximum transmit power of each relay can be set equal to  $1/(2N_r)$  of the average minimum total transmit power consumed in the entire network.

An interesting feature of the solution to the total power minimization problem is that this solution can be implemented in a distributed manner with minimal information exchange between the transceivers and the relays. This distributed implementation is presented in the following result.

### RESULT 9.13

In light of Eq. (9.50) and as the matrices  $\mathbf{D}_1$  and  $\mathbf{D}_2$  are diagonal, the optimal weight of the  $r$ th relay, denoted as  $w_r^o$ , is given by

$$w_r^o = \frac{\kappa \sqrt{\tilde{P}_1 (\gamma_1 + \gamma_2) f_{r1} f_{r2}}}{\sqrt{\lambda(\tilde{P}_1) ((\gamma_1 + \gamma_2) |f_{r2}|^2 + \lambda(\tilde{P}_1) (\tilde{P}_1 |f_{r1}|^2 + 1))}}. \quad (9.52)$$

*Continued*

**RESULT 9.13—CONT'D**

To implement this scheme, the relays do not require the knowledge of all channel coefficients, rather the  $r$ th relay node needs to know only its local channel coefficients  $f_{r1}$  and  $f_{r2}$  as well as three parameters, namely  $\tilde{P}_1$ ,  $\lambda(\tilde{P}_1)$ , and  $\kappa$ . These three parameters can be calculated and broadcasted by Transceiver 1 (which calculates  $\tilde{P}_1$ ) to all relays. Upon receiving these three parameters, the  $r$ th relay can use these three parameters along with the relay local CSI in Eq. (9.52) to calculate the relay amplification factor.

For the case of single-relay networks, the solution to the total power minimization problem is amenable to a closed-form solution, as presented in the following result [44].

**RESULT 9.14**

For the case of single-relay networks, the total power minimization problem (9.46) is amenable to a closed-form solution. Indeed, the optimal value of the single-relay coefficient, denoted as  $w$ , and the optimal values of the transceivers' transmit powers can be obtained as

$$\begin{aligned} w^o &= \sqrt{\frac{1}{|f_1 f_2| \sqrt{1 + (\gamma_1 + \gamma_2)^{-1}}}}, \\ P_1^o &= \frac{\sigma^2 \gamma_2 \left( |f_1| \sqrt{1 + (\gamma_1 + \gamma_2)^{-1}} + |f_2| \right)}{|f_1^2 f_2|}, \\ P_2^o &= \frac{\sigma^2 \gamma_1 \left( |f_2| \sqrt{1 + (\gamma_1 + \gamma_2)^{-1}} + |f_1| \right)}{|f_2^2 f_1|}, \end{aligned} \quad (9.53)$$

where  $f_1$  and  $f_2$  are the frequency-flat channel coefficients of the transceiver-relay links. In this case, the minimum total power is given by

$$P_T = \sigma^2 (\gamma_1 + \gamma_2) \left[ \frac{2 \sqrt{1 + (\gamma_1 + \gamma_2)^{-1}}}{|f_1 f_2|} + \frac{|f_1|^2 + |f_2|^2}{|f_1 f_2|^2} \right]. \quad (9.54)$$

It follows from Eq. (9.54) that if  $\gamma_1 + \gamma_2 \gg 1$  then

$$P_T = \sigma^2 (\gamma_1 + \gamma_2) \left( \frac{1}{|f_1|} + \frac{1}{|f_2|} \right)^2.$$

The solution to the single-relay case can be used to select, among several available relays, one relay which results in the lowest total transmit power consumption. To this end, one can choose the relay which has, among all relays, the largest harmonic mean of the magnitude of the relay local channel coefficients. It is noteworthy that at the optimum, when  $\gamma_1 = \gamma_2$  is chosen,  $P_1^o + P_2^o = 0.5 P_T$  holds true, and thus, the relay transmit power is half of the minimum total transmit power. That is, for  $\gamma_1 = \gamma_2$ ,

the minimum total transmit power is divided equally between the relay on one hand and the two transceivers on the other hand. This property allows the transmit power of the optimally selected relay to be controlled. One of the two transceivers can choose the power-optimal relay by calculating the harmonic means of the amplitudes of the channel coefficients of all relay, and choosing the relay which has the largest harmonic mean of the amplitudes of its channel coefficients. That transceiver can then broadcast the index of the selected relay to all relay nodes over a control channel. Upon “hearing” its own index over the control channel, the selected relay will use its local channel coefficients to obtain its own amplification factor  $w^o$  as in Eq. (9.53). Those relays, which do not detect their own index over the control channel, will not participate in the relaying scheme.

#### 9.4.1.2 Max-min SNR approach

In a max-min SNR fair design approach, the design of the network beamformer amounts to solving the optimization problem:

$$\max_{P_1, P_2, \mathbf{w}} \min(\text{SNR}_1(P_2, \mathbf{w}), \text{SNR}_2(P_1, \mathbf{w})), \text{ subject to } P_T \leq P_T^{\max}, P_1 \geq 0, P_2 \geq 0. \quad (9.55)$$

Here,  $P_T$ ,  $\text{SNR}_1$ , and  $\text{SNR}_2$ , are defined as in Eqs. (9.44) and (9.45), respectively, and  $P_T^{\max}$  is the maximum available total transmit power. Restricting the total power consumed in the entire network is valuable for network planning. Moreover, such a total transmit power constraint results in a guideline for choosing the maximum power consumption in individual relays. As will be shown later, when the max-min SNR approach is applied to the case of time-synchronous bidirectional two-way relay networks with  $N_r$  nodes, half of the maximum total transmit power is consumed in and is shared among the relays. In such a network, it is reasonable to assume that each relay, on average, consumes  $1/N_r$  of half of the total maximum power. This argument is in particular realistic when the relay nodes are moving randomly in the environment. In such a scenario, one can assume that different transceiver-relay channels are drawn from the same probability distribution. Another situation where a total power constraint is needed, is a scenario where all network nodes are powered up through the power grid (implying that they are stationary). In such a scenario, the total power consumed in the entire network has to be constrained. Note also that designing network beamforming schemes for bidirectional relay networks under per-node power constraints may not be amenable to a computationally efficient solution. As will be discussed later, with per-node power constraints, the max-min SNR optimal network beamforming schemes can have prohibitively high computational complexity [45]. Moreover, in bidirectional relay networks, for any given channel realization, the solution to the max-min SNR approach under per-node power constraints may not lead to an efficient design as it may happen that most of the relays do not transmit at their maximum allowable power [45]. As a result, the average power consumed by each relay for different realizations of the relay local channel will be less than the relay maximum available power. Because of all these reasons, total power constraints have been widely used in the literature for optimal

design and performance analysis of relay networks [12, 42, 46–49]. Note also that the max-min SNR design approach under a total power constraint results in an upper bound to the case where this approach is used under per-node power constraints. This is yet another reason the max-min SNR design approach under a total power constraint could be of interest.

To solve the optimization problem (9.55), it can be easily shown that at the optimum (i) the two transceivers' SNRs are balanced, i.e.,  $\text{SNR}_1 = \text{SNR}_2$  holds true, and (ii) the total power constraint in the optimization problem (9.55) is satisfied with equality. Based on these observations, the transceivers' transmit powers can be written, in terms of the beamforming weight vector  $\mathbf{w}$ , as

$$P_1 = \frac{P_T^{\max} - \sigma^2 \mathbf{w}^H \mathbf{w}}{2(1 + \mathbf{w}^H \mathbf{D}_1 \mathbf{w})} \geq 0, \quad P_2 = \frac{P_T^{\max} - \sigma^2 \mathbf{w}^H \mathbf{w}}{2(1 + \mathbf{w}^H \mathbf{D}_2 \mathbf{w})} \geq 0. \quad (9.56)$$

Based on Eq. (9.56), the max-min SNR optimization problem in Eq. (9.55) can be written as an unconstrained optimization problem given by

$$\max_{\mathbf{w}} \frac{(P_T^{\max}/\sigma^2 - \mathbf{w}^H \mathbf{w}) \mathbf{w}^H \mathbf{a} \mathbf{a}^H \mathbf{w}}{(1 + \mathbf{w}^H \mathbf{D}_1 \mathbf{w})(1 + \mathbf{w}^H \mathbf{D}_2 \mathbf{w})}, \quad \text{subject to } \mathbf{w}^H \mathbf{w} \leq P_T^{\max}/\sigma^2. \quad (9.57)$$

As proven in [42], the following result paves the way toward solving the optimization problem (9.57).

### RESULT 9.15

Any local maximum of the objective function of the optimization problem (9.57) is a global maximum. Also, the solution to the optimization problem (9.57), and therefore, the max-min-SNR optimal beamforming weight vector, is unique up to a scalar phase rotation.

It follows from Result 9.15 that a gradient-based algorithm can be used to obtain the optimal value of the network beamforming vector  $\mathbf{w}$ . However, a more computationally efficient technique is available as explained below [50].

### RESULT 9.16

The max-min-SNR optimal network beamforming vector  $\mathbf{w}$  can be obtained as

$$\mathbf{w}^0 = \bar{\kappa} \sqrt{2\tilde{P}_2} (2\tilde{P}_1 \mathbf{D}_1 + 2\tilde{P}_2 \mathbf{D}_2 + \mathbf{I}_{N_r})^{-1} \mathbf{a}, \quad (9.58)$$

where  $\tilde{P}_2 = 0.5P_T^{\max}/\sigma^2 - \tilde{P}_1$  and the parameter  $\kappa$  is obtained as

$$\bar{\kappa} \triangleq \left( \mathbf{a}^H (\mathbf{I}_{N_r} + 2\tilde{P}_1 \mathbf{D}_1) (2\tilde{P}_1 \mathbf{D}_1 + 2\tilde{P}_2 \mathbf{D}_2 + \mathbf{I}_{N_r})^{-2} \mathbf{a} \right)^{-1/2}. \quad (9.59)$$

And the parameter  $\tilde{P}_1$  is obtained as the *provably unique* solution to the following equation:

$$(P_T^{\max}/\sigma^2 - 4\tilde{P}_1) \mathbf{a}^H (2\tilde{P}_1 \mathbf{D}_1 + \mathbf{I}_{N_r} + (P_T^{\max}/\sigma^2 - 2\tilde{P}_1) \mathbf{D}_2)^{-1} \mathbf{a} \\ - \tilde{P}_1 (P_T^{\max}/\sigma^2 - 2\tilde{P}_1) \mathbf{a}^H (2\tilde{P}_1 \mathbf{D}_1 + \mathbf{I}_{N_r} + (P_T^{\max}/\sigma^2 - 2\tilde{P}_1) \mathbf{D}_2)^{-2} (2\mathbf{D}_1 - 2\mathbf{D}_2) \mathbf{a} = 0. \quad (9.60)$$

To solve Eq. (9.60), one can use either a Newton-Raphson method or a simple bisection technique. In fact, the LHS of Eq. (9.60) is positive/negative for those values of  $\mu_n$  which are smaller/larger than the unique solution to Eq. (9.60). The optimal values of the transceivers' transmit powers are then obtained from Eq. (9.56), when  $\mathbf{w}$  is replaced  $\mathbf{w}^o$  given in Eq. (9.56) or as  $P_1^o = \sigma^2 \tilde{P}_1$  and  $P_2^o = \sigma^2 \tilde{P}_2$ .

The maximum balanced SNR is given as

$$\text{SNR}_{\max} = 2\tilde{P}_1\tilde{P}_2 \quad \mathbf{a}^H (2\tilde{P}_1\mathbf{D}_1 + 2\tilde{P}_2\mathbf{D}_2 + \mathbf{I}_{N_r})^{-1} \mathbf{a}. \quad (9.61)$$

For the case of single-relay networks, the solution to the max-min SNR problem is amenable to a closed-form solution, as presented in the following result [51].

### RESULT 9.17

For single-relay networks, the solution to the max-min SNR design problem, under a total power constraint, is amenable to a closed-form solution given as

$$w^o = \sqrt{\frac{P_T^{\max}}{\sigma^2 + \sqrt{(\sigma^2 + P_T^{\max}|f_1|^2)(\sigma^2 + P_T^{\max}|f_2|^2)}}}. \quad (9.62)$$

Here,  $w^o$  is the optimal value of the real-valued relay weight, while  $f_1$  and  $f_2$  are the frequency-flat channel coefficients of the transceiver-relay links. Substituting Eq. (9.62) in Eq. (9.48), the optimal values of  $P_1$  and  $P_2$  are respectively given as

$$P_1^o = \frac{P_T^{\max} \sqrt{\sigma^2 + P_T^{\max}|f_2|^2}}{2\sqrt{\sigma^2 + P_T^{\max}|f_1|^2} + 2\sqrt{1 + P_T^{\max}|f_2|^2}}, \quad P_2^o = \frac{P_T^{\max} \sqrt{\sigma^2 + P_T^{\max}|f_1|^2}}{2\sqrt{\sigma^2 + P_T^{\max}|f_1|^2} + 2\sqrt{\sigma^2 + P_T^{\max}|f_2|^2}}, \quad (9.63)$$

and the maximum balanced SNR is given by

$$\text{SNR}_1 = \text{SNR}_2 = \frac{(P_T^{\max})^2 |f_1 f_2|^2}{2\sigma^2 \left( \sqrt{\sigma^2 + P_T^{\max}|f_1|^2} + \sqrt{\sigma^2 + P_T^{\max}|f_2|^2} \right)^2}. \quad (9.64)$$

Note that if  $P_T^{\max}/\sigma^2 \gg \max(1/|f_1|^2, 1/|f_2|^2)$ , then

$$\text{SNR}_1 = \text{SNR}_2 \approx \frac{|f_1 f_2|^2 P_T^{\max}}{2\sigma^2 (|f_1| + |f_2|)^2} = \frac{P_T^{\max}}{2\sigma^2 \left( \frac{1}{|f_1|} + \frac{1}{|f_2|} \right)^2}. \quad (9.65)$$

The solution to the max-min SNR design approach for the single-relay case can be used to select, among several available relays, one relay which results in the largest balanced SNR under a given total power budget. Indeed, it follows from Eq. (9.65) that one can select the relay which has, among all relays, the largest harmonic mean of the magnitude of the relay local channel coefficients.

An interesting feature of the SNR balancing solution is that this solution can be implemented in a distributed manner as explained in Result 9.18 [50].

**RESULT 9.18**

Since the matrix  $(2\tilde{P}_1\mathbf{D}_1 + 2\tilde{P}_2\mathbf{D}_2 + \mathbf{I})^{-1}$  in Eq. (9.58) is diagonal, the  $r$ th relay can obtain its own beamforming weight  $w_r^0$  using its own local channel information along with the knowledge of only two additional parameters, namely  $\bar{\kappa}$  and  $\tilde{P}_1$ , that are common for all the relays. Therefore, if one of the two transceivers broadcasts parameters  $\bar{\kappa}$  and  $\tilde{P}_1$ , the  $r$ th relay can use these two parameters along with its own local channel information (i.e.,  $f_{r1}$  and  $f_{r2}$ ) to obtain its optimal beamforming weight as

$$w_r^0 = \bar{\kappa} \sqrt{2\tilde{P}_2} (2\tilde{P}_1|f_{r1}|^2 + 2\tilde{P}_2|f_{r2}|^2 + 1)^{-1} f_{r1} f_{r2} \quad (9.66)$$

where we have used the fact that the  $r$ th entry of vector  $\mathbf{a} = \mathbf{f}_1 \odot \mathbf{f}_2$  is equal to  $f_{r1} f_{r2}$ . Note that each relay can obtain  $\tilde{P}_2$  as  $\tilde{P}_2 = 0.5P_T^{\max}/\sigma^2 - \tilde{P}_1$ .

**9.4.1.3 Sum-rate maximization**

Maximizing the sum-rate subject to a total power constraint is yet another approach to design a network beamformer for bidirectional relay networks. In this approach, the design parameters are obtained such that the sum-rate is maximized under a total power constraint. That is this approach amounts to solving the following optimization problem:

$$\max_{\mathbf{p} \geq 0, \mathbf{w}} \quad R_1(P_2, \mathbf{w}) + R_2(P_1, \mathbf{w}), \quad \text{subject to} \quad P_T \leq P_T^{\max}, \quad (9.67)$$

where the following definitions are used:  $\mathbf{p} \triangleq [P_1 \ P_2]^T$ ,  $R_1(P_2, \mathbf{w}) \triangleq \frac{1}{2} \log_2(1 + \text{SNR}_1(P_2, \mathbf{w}))$ ,  $R_2(P_1, \mathbf{w}) \triangleq \frac{1}{2} \log_2(1 + \text{SNR}_2(P_1, \mathbf{w}))$ . Since  $R_1(P_2, \mathbf{w}) + R_2(P_1, \mathbf{w}) = \frac{1}{2} \log_2((1 + \text{SNR}_1(P_2, \mathbf{w}))(1 + \text{SNR}_2(P_1, \mathbf{w})))$ , the sum-rate maximization problem (9.67) is equivalent to the following optimization problem:

$$\max_{\mathbf{p} \geq 0, \mathbf{w}} \quad (1 + \text{SNR}_1(P_2, \mathbf{w}))(1 + \text{SNR}_2(P_1, \mathbf{w})), \quad \text{subject to} \quad P_T \leq P_T^{\max}. \quad (9.68)$$

As proven in [52], the following result paves the way toward solving the optimization problem (9.68).

**RESULT 9.19**

The achievable SNR region is described as

$$\text{SNR}_1(P_2, \mathbf{w}) + \text{SNR}_2(P_1, \mathbf{w}) \leq 2\gamma_{\max}(P_T^{\max}), \quad (9.69)$$

where  $\gamma_{\max}(P_T^{\max})$  is the maximum achievable balanced SNR obtained under a total power budget, that is,  $\gamma_{\max}(P_T^{\max})$  is given by the optimal value of the objective function in the optimization problem (9.55).

This result is quite useful to obtain the sum-rate-optimal network beamforming weight vector and the corresponding power allocation scheme. One can easily show that at the optimum of the maximization problem (9.68), the SNR pair must be on the boundary of the SNR region. Otherwise, we can find a higher  $\text{SNR}_1(P_2, \mathbf{w})$  with  $\text{SNR}_2(P_1, \mathbf{w})$  fixed (or vice versa), which results in a higher value of the objective

function in the optimization problem (9.68). Hence, without loss of optimality, the constraint  $\text{SNR}_1(P_1, \mathbf{w}) + \text{SNR}_2(P_1, \mathbf{w}) = 2\gamma_{\max}(P_T^{\max})$  can be added to the sum-rate maximization optimization problem in the optimization problem (9.68) as

$$\begin{aligned} & \max_{\mathbf{p} \geq 0, \mathbf{w}} (1 + \text{SNR}_1(P_2, \mathbf{w})) (1 + \text{SNR}_2(P_1, \mathbf{w})), \\ & \text{subject to } P_T \leq P, \text{SNR}_1(P_2, \mathbf{w}) + \text{SNR}_2(P_1, \mathbf{w}) = 2\gamma_{\max}(P_T^{\max}). \end{aligned} \quad (9.70)$$

Based on the fact that

$$(1 + \text{SNR}_1(P_2, \mathbf{w})) (1 + \text{SNR}_2(P_1, \mathbf{w})) \leq \left(1 + \frac{\text{SNR}_1(P_2, \mathbf{w}) + \text{SNR}_2(P_1, \mathbf{w})}{2}\right)^2, \text{ with}$$

the equality being satisfied only when  $\text{SNR}_1(P_2, \mathbf{w}) = \text{SNR}_2(P_1, \mathbf{w})$  holds true, one concludes that at the optimum of the optimization problem (9.68),  $\text{SNR}_1(P_2, \mathbf{w}) = \text{SNR}_2(P_1, \mathbf{w}) = \gamma_{\max}(P_T^{\max})$  holds true, leading to the following conclusion:

### RESULT 9.20

Under a total power constraint, the sum-rate maximization approach to the design of jointly optimal network beamformer and power allocation leads to the same solution as the max-min SNR fair design approach does.

It follows from the above results that Eqs. (9.58)–(9.60) can be used to obtain the sum-rate-optimal network beamforming and the corresponding power allocation scheme.

It is worth mentioning that in parallel channels, sum-rate-optimal power allocation often leads to a water-filling type of solution where the channels with higher SNRs receive larger portions of the total available power as compared to channels with smaller SNRs. Such solutions are often different from solutions obtained via a max-min fair design approach, where user fairness is the primary objective. Interestingly, for the bidirectional network beamforming problem, unlike traditional water-filling schemes, the sum-rate-optimal power allocation under a total power constraint leads to an SNR balancing (or max-min SNR fair) design solution. As a result, *the transceiver with stronger links to the relays consumes less transmit power as compared to the transceiver with weaker channels to the relays*.

#### 9.4.1.4 Individual power constraints

The authors of [45] consider the design of optimal joint power control and network beamforming for two-way relay networks under per-node power constraints. Using a max-min SNR fair design approach, this study presents an analytical solution for single-relay networks. For multi-relay networks, the study in [45] proposes a numerical solution to the max-min SNR fair design approach under per-node power constraints. This solution is composed of a combination of a two-dimensional search

over the feasible values of the transceivers' transmit powers and a second order convex cone feasibility programming. Indeed, in this solution, the plane of feasible values of  $(P_1, P_2)$  is first discretized into a sufficiently fine grid, and then, for each vertex on this grid, a second order convex cone feasibility programming approach is solved. In this approach, the smaller of the two transceivers' SNRs is maximized over the relay beamforming vector  $\mathbf{w}$  under per-relay powers constraints. The largest max-min SNR achieved over the  $(P_1, P_2)$  grid yields the optimal values of  $P_1$  and  $P_2$ , and eventually, the optimal value of  $\mathbf{w}$ . Unfortunately, the computational complexity of this solution can be prohibitively high especially for networks with many relays. The investigation in [45] also presents two computationally affordable suboptimal solutions, which iteratively optimize the transceivers' powers and the relay beamformer weight vector. Simulation results of [45] show that these suboptimal solutions can perform close to the optimal solution.

#### **9.4.1.5 TDBC versus MABC**

The investigation in [41] compares the performance of two bidirectional network beamforming schemes, namely the MABC strategy and the TDBC protocol, under joint optimal power control and beamforming design. To do so, this study first designs two TDBC-based bidirectional network beamformers, through minimizing the total power consumed in the entire network subject to QoS constraints, for the two cases with and without a direct link between the two transceivers. These solutions are then used to compare the performance of the underlying TDBC-based approach to that of the MABC-based technique. The importance of this comparison resides in the fact that the TDBC approach appears to have certain advantages compared to the MABC protocol. These advantages are twofold: (1) the TDBC-based solution benefits from certain additional degrees of freedom compared to the MABC method, and (2) the TDBC method can also benefit from the availability of a direct link between the two transceivers, while the MABC approach cannot exploit this direct link due to the fact that the relays are half-duplex. Interestingly, it is shown in [41] that in the absence of a direct link between the two transceivers, when the QoS constraints are imposed such that given probabilities of uncoded error (or equivalently, to meet given SNR constraints) are satisfied, both TDBC- and MABC-based schemes perform closely in terms of the minimum total transmit power. However, when the QoS constraints are used to ensure that given rates are achieved, the MABC-based scheme outperforms the TDBC counterpart. When a direct link exists between the two transceivers, the story is different: the TDBC-based protocol can outperform the MABC-based strategy even for rate satisfying QoS constraints, provided that the direct link is strong enough.

#### **9.4.2 ASYNCHRONOUS NETWORKS**

In the previous subsection, the focus was on synchronous relay networks, where the delays of signal propagation from/to each transceiver to/from different relays are assumed to be identical (or within one symbol period). This assumption

implies that the signals transmitted by the transceivers arrive at the relays with the same delay and the signals transmitted by different delays also arrive at each transceiver with the same delay. Such an assumption may not be realistic when transceiver-relay distances are significantly different from each other. For example, in an LTE network with a symbol rate of 20 Ms/s, if the distances from one transceiver to two relays are different from each other by only 15 meters, the symbol stream transmitted by this transceiver will arrive at the two relays with one symbol differential delay. Similarly, with only 15-meter difference between the distances from two relays to one transceiver, the signals transmitted by the two relays will arrive at the transceiver with one symbol differential delay. As a result, the end-to-end channel will no longer be frequency-flat, rather a time-dispersive (frequency-selective) model appears to be more realistic. Indeed, the end-to-end channel can be viewed as a multi-path link which could result in ISI. Hence, for such an ISI channel, equalization becomes inevitable.

Note however that there is an important distinction between traditional models of multi-path channels and an asynchronous relay channel. This distinction resides in the fact that in traditional multi-path channel models, the end-to-end CIR is often out of the control of the system designer and the channel has to be compensated for at the transmitter, through precoding, or at the receiver, through channel equalization. In a relay channel, however, the end-to-end CIR can somewhat be adjusted somewhere between the two end nodes (i.e., at the relays) for optimal performance. In other words, in asynchronous one-way or two-way relay channels, the problem of optimal design of transmit strategies is intertwined with *channel design*—a concept which does not appear when dealing with traditional multi-path channel models. This concept brings new dimensions to the design of the end-to-end link leading to non-intuitive and interesting results.

To combat the ISI produced due to asynchronism in relay networks with frequency-flat transceiver-relay channels or that caused by the frequency selectivity of the transceiver-relay links in such networks, three different approaches have been studied in the literature. One approach, referred to as FF relying scheme, suggests that the relays be equipped with FIR filters while transceivers may or may not use FIR filters at their receiver front-ends. In the FF scheme, the equalization of the end-to-end channel is implemented in a distributed method. The second approach to tackle ISI at the two transceivers relies on employing multicarrier transmission and reception, i.e., orthogonal frequency-division multiplexing (OFDM), schemes *in all nodes* to “diagonalize” the end-to-end channel followed by AF relaying over each subcarrier. Both approaches rely on some level of processing complexity at the relays, let it be equipping the relay nodes with FIR filters or furnishing these nodes with OFDM decoder and encoder.

Another line of published results relaxes the relay processing to simple AF relaying and leaves the burden of the channel equalization to the two transceivers—either through precoding the transmitter front-ends of the transceivers (i.e., pre-channel equalization) or through channel equalization at the receiver front-ends of the two transceivers (i.e., post-channel equalization) or through joint pre-channel and post-channel equalization.

In the rest of this subsection, first the end-to-end channel model is revisited and new notations are introduced. Then, assuming different optimality criteria, the designs of network beamforming are presented for both multi-carrier and single-carrier networks under a total power budget. Also, reviewed is the total power minimization approach to design network beamforming schemes under QoS constraints.

#### 9.4.2.1 End-to-end channel model

For a two-way relay channel with frequency-flat reciprocal transceiver-relay channels, the end-to-end CIR in Eq. (9.1) can be written as<sup>3</sup>

$$h[n] \triangleq h_{12}[n] = h_{21}[n] = \sum_{r=1}^{N_r} f_{r1} f_{r2} w_r \delta[n - (\check{n}_{r1} + \check{n}_{r2})]. \quad (9.71)$$

In this model, each relay contributes only to one of the taps of the end-to-end CIR. That is, the  $r$ th relay contributes only to the tap  $\check{n}_r \triangleq \check{n}_{r1} + \check{n}_{r2}$ . Note also that if  $\check{n}_{r_1} = \check{n}_{r_2}$  holds true for some  $r_1 \neq r_2$ , then the  $r_1$ th and the  $r_2$ th relays contribute to the same tap. It follows from Eq. (9.71) that the vector of taps of the end-to-end CIR can be written as

$$\mathbf{h}(\mathbf{w}) = \mathbf{Aw}, \quad (9.72)$$

where  $\mathbf{h}(\mathbf{w}) \triangleq [h[0] \ h[1] \ \dots \ h[N_c - 1]]^T$ ,  $N_c$  is the *maximum possible length* (i.e., maximum possible delay spread) of the end-to-end CIR, i.e.,  $N_c = 1 + \max_{r \in \mathcal{N}_r} (\check{n}_{r1} + \check{n}_{r2})$ ,  $\mathcal{N}_r \triangleq \{1, 2, \dots, N_r\}$ , and  $\mathbf{A}$  is an  $N_c \times N_r$  matrix, whose  $(n+1, r)$ th entry is defined as

$$A(n+1, r) = \begin{cases} f_{r1} f_{r2}, & (n-1)T_s < \tau_r \leq nT_s, \\ 0, & \text{otherwise.} \end{cases} \quad \text{for } n = 0, 1, \dots, N_c - 1, \text{ and } r = 1, 2, \dots, N_r. \quad (9.73)$$

Here,  $\check{n}_{r1} + \check{n}_{r2} = \lceil \tau_r / T_s \rceil$  is the discrete-time delay of the end-to-end relaying path which goes through the  $r$ th relay while  $\tau_r$  is the corresponding continuous-time delay. Indeed,  $A(n+1, r)w_r$  describes the contribution of the  $r$ th relaying path to the  $n$ th tap of  $h[\cdot]$ , for  $n = 0, 1, \dots, N_c - 1$  and  $r = 1, 2, \dots, N_r$ .

Note that for a synchronous network, matrix  $\mathbf{A}$  has only one nonzero row. For an asynchronous network, however, matrix  $\mathbf{A}$  can have as many as  $N_r$  nonzero rows. Indeed, when each relay contributes to one distinct tap of the end-to-end CIR, this CIR has  $N_r$  nonzero taps which are spread from tap 0 to tap  $N_c - 1$ .

---

<sup>3</sup>Recall that the reciprocity of the transceiver-relay links implies that  $\check{f}_{qr}[\cdot] = \check{g}_{rq}[\cdot]$  holds true for  $q = 1, 2$ , and  $r = 1, 2, \dots, N_r$ . Recall also that the frequency-flatness of the transceiver-relay links means that the CIRs of the transceiver-relay links can be written as  $\check{f}_{qr}[n] = f_{rq}\delta[n - \check{n}_{rq}]$ , where  $\check{n}_{rq}$  represents the time delay (in terms of the number of symbols) of the link between the  $r$ th relay and Transceiver  $q$  and  $f_{rq}$  is the corresponding flat fading channel coefficient, for  $q = 1, 2$  and  $r = 1, 2, \dots, N_r$ .

#### 9.4.2.2 Multi-carrier equalization

Furnishing the two transceivers with OFDM transmission and reception schemes, while the relays rely on simple AF relaying scheme, belongs to the class of joint pre-channel and post-channel equalization and network beamforming schemes. In this approach, the pre-channel equalizer (precoding) matrix used at the front-end of each transmitter is the inverse discrete Fourier transform (DFT) matrix and the post-channel equalizer matrix is the DFT matrix. This approach assumes a block communication scheme and relies on cyclic prefix (CP) insertion to avoid inter-block-interference [54].

Fig. 9.4 shows this scheme: at each transceiver, the stream of the information symbols first goes through a serial-to-parallel conversion operation, denoted as “S/P” block, which produces blocks of symbols. The symbol blocks then go through a power allocation block followed by inverse DFT operation (i.e., they are multiplied by the inverse DFT matrix  $\mathbf{F}^H$ ). The output blocks of the DFT operation are appended with CP, and then, are converted into serial signals using parallel-to-serial conversion operation, denoted as “P/S” block. The serial signals of the two transceivers are then simultaneously transmitted over the asynchronous relay channel. At the receiver front-end of each transceiver, using an “S/P” block, the received noisy signal is first converted into blocks of signals. The CP is then removed from the received signal blocks. The CP-free signal blocks go through the DFT block (i.e., they are multiplied by the DFT matrix  $\mathbf{F}$ ) followed by a “P/S” block. After self-interference cancelation, the so-obtained serial signals are processed to detect the transmitted symbols of interest.

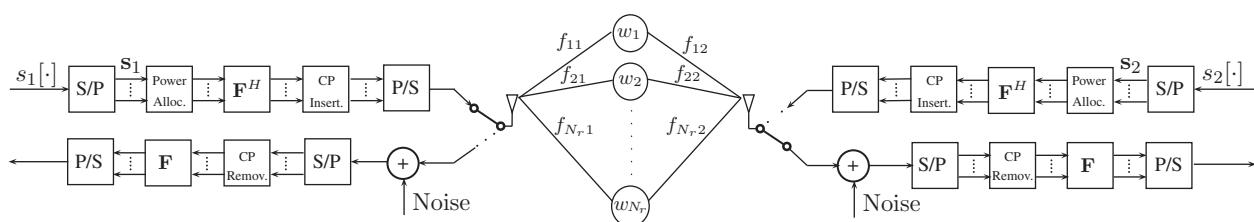
In the next subsections, two design approaches will be presented: (1) a max-min SNR fair design approach under a total power budget for the entire network, and (2) a sum-rate maximization technique also under a total power budget for the entire network.

#### 9.4.2.3 Max-min SNR fair design approach

Consider a two-way relaying scheme, where the end-to-end CIR is compensated for (is diagonalized) by employing OFDM transmission and reception schemes at the two transceivers, while the relays resort to simple AF relaying scheme. In such a network, the design parameters are the transceivers' transmit powers over different subcarriers and the relay beamforming weight vector. In a max-min SNR fair design approach under a total power constraint, the joint design of network beamforming and power allocation amounts to solving the following optimization problem:

$$\max_{\mathbf{p}_1, \mathbf{p}_2 \geq 0} \quad \max_{\mathbf{w}} \min_{i \in \mathcal{N}_f} \min_{q \in \{1, 2\}} \text{SNR}_{iq}(P_{i\bar{q}}, \mathbf{w}), \quad \text{subject to} \quad \frac{\mathbf{1}^T \mathbf{p}_1}{N_f} + \frac{\mathbf{1}^T \mathbf{p}_2}{N_f} + \sum_{r=1}^{N_r} Q_r \leq P_T^{\max}, \quad (9.74)$$

where  $\bar{q} = 2$  if  $q = 1$ , and  $\bar{q} = 1$  when  $q = 2$ ;  $\text{SNR}_{iq}(P_{i\bar{q}}, \mathbf{w})$  is the received SNR at the  $q$ th transceiver over the  $i$ th subcarrier, for  $q = 1, 2$  and  $i \in \mathcal{N}_f \triangleq \{1, 2, \dots, N_f\}$ ;  $\mathbf{p}_q \triangleq [P_{1q} \ P_{2q} \ \dots \ P_{N_f q}]^T$  is the vector of the transmit powers of Transceiver  $q$  over all subcarriers;  $\mathbf{1}$  is the vector of all ones;  $P_{iq}$  is the transmit power of the  $q$ th

**FIG. 9.4**

Block diagram of an OFDM-based two-way relay network [53].

transceiver over the  $i$ th subcarrier; and  $Q_r$  is the transmit power of the  $r$ th relay. Note also that  $\frac{\mathbf{1}^T \mathbf{p}_1}{N_f}$  and  $\frac{\mathbf{1}^T \mathbf{p}_2}{N_f}$  are the transmit powers of Transceivers 1 and 2, respectively, while the total relay transmit power is given by

$$\sum_{r=1}^{N_r} Q_r = \frac{\mathbf{1}^T \mathbf{p}_1}{N_f} (\mathbf{w}^H \mathbf{D}_1 \mathbf{w}) + \frac{\mathbf{1}^T \mathbf{p}_2}{N_f} (\mathbf{w}^H \mathbf{D}_2 \mathbf{w}) + \sigma^2 \mathbf{w}^H \mathbf{w}. \quad (9.75)$$

Here, as defined earlier,  $\mathbf{D}_1 \triangleq \text{diag}(\mathbf{f}_1 \odot \mathbf{f}_1^*)$  and  $\mathbf{D}_2 \triangleq \text{diag}(\mathbf{f}_2 \odot \mathbf{f}_2^*)$  are two  $N_r \times N_r$  diagonal matrices. The  $N_f \times 1$  vector of the signals received at Transceiver  $q$  over all  $N_f$  subcarriers, denoted as  $\mathbf{r}_q$ , can be expressed as

$$\mathbf{r}_q = \text{diag}\left(\left\{\sqrt{N_f} \mathbf{f}_i^H \tilde{\mathbf{h}}(\mathbf{w})\right\}_{i=1}^{N_f}\right) \text{diag}(\mathbf{p}_{\bar{q}}) \mathbf{s}_{\bar{q}} + \text{received noise}, \quad (9.76)$$

where  $\mathbf{s}_{\bar{q}}$  is the  $N_f \times 1$  vector of the symbols transmitted by Transceiver  $\bar{q}$  over the  $N_f$  subcarriers,  $\mathbf{f}_i \triangleq \frac{1}{\sqrt{N_f}} \left[ 1 \ e\left(j\frac{2\pi(i-1)}{N_f}\right) \ e\left(j\frac{4\pi(i-1)}{N_f}\right) \dots \ e\left(j\frac{2(N_f-1)(i-1)\pi}{N_f}\right) \right]^T$  is the  $i$ th Vandermonde column vector of  $\mathbf{F}^H$ ,  $\mathbf{F}$  is the  $N_f \times N_f$  DFT matrix,  $\tilde{\mathbf{h}}$  is the zero-padded version of the vector of end-to-end CIR taps  $\mathbf{h}$ , that is,  $\tilde{\mathbf{h}}(\mathbf{w}) \triangleq \left[ \mathbf{h}^T(\mathbf{w}) \ \mathbf{0}_{1 \times (N_f - N_c)} \right]^T$ . Based on these definitions, in Eq. (9.76), the expression  $\sqrt{N_f} \mathbf{f}_i^H \tilde{\mathbf{h}}(\mathbf{w})$  is the frequency response of the end-to-end CIR at the  $i$ th subcarrier. Also the noise term in Eq. (9.76) is the sum of two noise vectors: the vector of subcarrier noises at Transceiver  $q$  and a vector which depends on the relay noises. As shown in Fig. 9.4, this is how the latter noise vector is formed: at each relay, the received noise is multiplied by the relay beamforming weight. This weighted noise then goes through the channel between that relay and Transceiver  $\bar{q}$ , and thus, is delayed and attenuated accordingly. The delayed and attenuated versions of weighted noises of different relays add up at Transceiver  $\bar{q}$ . Then, this relay noise mixture received at Transceiver  $\bar{q}$  goes through an “S/P” block (i.e., it is turned into a vector), and then, passes through the CP removal block followed by the DFT operation.

Based on Eq. (9.76), it can be shown that

$$\text{SNR}_{iq}(P_{\bar{q}}, \mathbf{w}) = \frac{N_f P_{\bar{q}} |\mathbf{f}_i^H \tilde{\mathbf{h}}(\mathbf{w})|^2}{\sigma^2 (\mathbf{w}^H \mathbf{D}_q \mathbf{w} + 1)}. \quad (9.77)$$

It is worth mentioning that in Eq. (9.77), the term  $N_f |\mathbf{f}_i^H \tilde{\mathbf{h}}(\mathbf{w})|^2$  is the squared magnitude of the frequency response of the end-to-end CIR at the  $i$ th subcarrier and the term  $\sigma^2 \mathbf{w}^H \mathbf{D}_q \mathbf{w}$  is the power of the relay noises after relay amplification and when these noises add up at Transceiver  $q$ .

As proven in [53], solving the max-min problem (9.74) relies on the following SNR balancing result.

### RESULT 9.21

At the optimum of the max-min problem (9.74), the SNRs over all subcarriers at both transceivers are equal.

Based on this result and using the total power constraint, one can show that at the optimum, the transceivers' subcarrier powers can be written, in terms of  $\mathbf{w}$ , as

$$P_{iq} = \frac{N_f(P_T^{\max} - \sigma^2 \mathbf{w}^H \mathbf{w})}{2(\mathbf{w}^H \mathbf{D}_q \mathbf{w} + 1) |\mathbf{f}_i^H \tilde{\mathbf{h}}(\mathbf{w})|^2 \mathbf{1}^T \mathbf{u}(\mathbf{w})}, \text{ for } i = 1, 2, \dots, N_f, \text{ and } q = 1, 2. \quad (9.78)$$

Here, the  $N_f \times 1$  vector  $\mathbf{u}(\mathbf{w})$  is defined as  $\mathbf{u}(\mathbf{w}) \triangleq \left[ \frac{1}{|\mathbf{f}_1^H \tilde{\mathbf{h}}(\mathbf{w})|^2} \frac{1}{|\mathbf{f}_2^H \tilde{\mathbf{h}}(\mathbf{w})|^2}, \dots, \frac{1}{|\mathbf{f}_{N_f}^H \tilde{\mathbf{h}}(\mathbf{w})|^2} \right]^T$ . Using Eq. (9.78), one can show that the optimization problem (9.74) is equivalent to the following maximization:

$$\max_{\mathbf{w}} \frac{N_f^2 (P_T^{\max} / \sigma^2 - \mathbf{w}^H \mathbf{w})}{2[(\mathbf{w}^H \mathbf{D}_1 \mathbf{w} + 1)(\mathbf{w}^H \mathbf{D}_2 \mathbf{w} + 1)] \mathbf{1}^T \mathbf{u}(\mathbf{w})} \text{ subject to } \mathbf{w}^H \mathbf{w} \leq \frac{P_T^{\max}}{\sigma^2}, \quad (9.79)$$

The following result, which is proven in [53], facilitates further simplification of the maximization problem (9.79).

### RESULT 9.22

The solution to the optimization problem (9.79) is such that only one of the entries of the vector of the taps of the end-to-end CIR  $\mathbf{h}(\mathbf{w}) = \mathbf{A}\mathbf{w}$  is nonzero.

**Result 9.22** states that at the optimum of the optimization problem (9.74), only one of the taps of the end-to-end CIR is nonzero. Thus, in order to achieve the maximum balanced SNR, all relays, which contribute to the zero taps of the end-to-end CIR, should be assigned a zero weight, i.e., those relays should be turned off, and only the relays which contribute to the only nonzero tap of the end-to-end CIR should be turned on. Let  $\mathcal{U}_n$  represent the set of those values of  $\mathbf{w}$  which have zero entries for those relays which do not contribute to the  $n$ th tap of the end-to-end CIR and could have nonzero entries for those relays which do contribute to the  $n$ th tap of the end-to-end CIR. Note that  $\mathcal{U}_n \cap \mathcal{U}_m = \emptyset$ , for  $m \neq n$ , as no relay can contribute to two taps of the end-to-end CIR. **Result 9.22** implies that at the optimum of the optimization problem (9.74),  $\mathbf{w} \in \bigcup_{n=0}^{N_c-1} \mathcal{U}_n$  holds true, and thus, all entries of the vector  $\mathbf{u}(\mathbf{w})$  will be identical (e.g., they are all equal to  $\frac{1}{|\mathbf{f}_1^H \tilde{\mathbf{h}}(\mathbf{w})|^2}$ ). Hence, for  $\mathbf{w} \in \bigcup_{n=0}^{N_c-1} \mathcal{U}_n$ , one can write

$$\begin{aligned} \mathbf{1}^T \mathbf{u}(\mathbf{w}) &= \left( |\mathbf{f}_1^H \tilde{\mathbf{h}}(\mathbf{w})|^2 \right)^{-1} \mathbf{1}^T \mathbf{1} = \left( \frac{1}{N_f} \sum_{i=1}^{N_f} |\mathbf{f}_i^H \tilde{\mathbf{h}}(\mathbf{w})|^2 \right)^{-1} N_f \\ &= N_f^2 (\tilde{\mathbf{h}}^H(\mathbf{w}) \tilde{\mathbf{h}}(\mathbf{w}))^{-1} = N_f^2 (\mathbf{h}^H(\mathbf{w}) \mathbf{h}(\mathbf{w}))^{-1} = N_f^2 (\mathbf{w}^H \mathbf{A}^H \mathbf{A} \mathbf{w})^{-1}, \end{aligned} \quad (9.80)$$

where Parseval's theorem is used in the third equality, the fact that  $\tilde{\mathbf{h}}(\mathbf{w})$  is a zero-padded version of  $\mathbf{h}(\mathbf{w})$  is used in the fourth equality, and Eq. (9.72) is used in the fifth equality.

With Eq. (9.80) and based on the above discussion, the optimization problem (9.79) can be written as

$$\max_{\mathbf{w}} \frac{(P_T^{\max}/\sigma^2 - \mathbf{w}^H \mathbf{w}) \mathbf{w}^H \mathbf{A}^H \mathbf{A} \mathbf{w}}{2[(\mathbf{w}^H \mathbf{D}_1 \mathbf{w} + 1)(\mathbf{w}^H \mathbf{D}_2 \mathbf{w} + 1)]}, \text{ subject to } \mathbf{w}^H \mathbf{w} \leq \frac{P_T^{\max}}{\sigma^2} \text{ and } \mathbf{w} \in \bigcup_{n=0}^{N_c-1} \mathcal{U}_n, \quad (9.81)$$

or equivalently as

$$\max_{n \in \mathcal{N}_c} \max_{\mathbf{w} \in \mathcal{U}_n} \frac{(P_T^{\max}/\sigma^2 - \mathbf{w}^H \mathbf{w}) \mathbf{w}^H \mathbf{A}^H \mathbf{A} \mathbf{w}}{2[(\mathbf{w}^H \mathbf{D}_1 \mathbf{w} + 1)(\mathbf{w}^H \mathbf{D}_2 \mathbf{w} + 1)]}, \text{ subject to } \mathbf{w}^H \mathbf{w} \leq \frac{P_T^{\max}}{\sigma^2}, \quad (9.82)$$

where the following definition is used:  $\mathcal{N}_c \triangleq \{1, 2, \dots, N_c\}$ . The optimization problem (9.82) follows from the optimization problem (9.81) based on the fact that the sets  $\{\mathcal{U}_n\}_{n=0}^{N_c-1}$  are mutually exclusive. In the optimization problem (9.82), the inner maximization aims to find the value of  $\mathbf{w}$  which maximizes the balanced SNR, assuming that only the  $n$ th tap of the end-to-end CIR is nonzero, for any given  $n$ . The outer maximization finds the optimal index of the nonzero tap which results in the largest balanced SNR. It is worth mentioning that when only the relays which contribute to one tap of the end-to-end CIR participate in relaying, the end-to-end channel will become frequency-flat, i.e.,  $|f_i^H \tilde{\mathbf{h}}(\mathbf{w})| = |f_j^H \tilde{\mathbf{h}}(\mathbf{w})|$  holds for the optimal  $\mathbf{w}$ , for any  $i, j \in \{1, 2, \dots, N_f\}$ . In other words, activating only one set of such relays will render the network synchronous. Thus, an SNR balancing problem can be solved for each set of the relays, which contribute to one of the taps of the end-to-end CIR  $h[\cdot]$ , with the aim to calculate the corresponding maximum balanced subcarrier SNR. The so-obtained balanced subcarrier SNRs for different taps are then examined to find the largest balanced SNR among these SNRs and the corresponding set of the relays (which have to be selected to participate in relaying while the rest of the relays have to be turned off). This is exactly what solving the optimization (9.82) means.

To solve the optimization problem (9.82), note that if  $\mathbf{w} \in \mathcal{U}_n$ , for  $n = 0, 1, N_c - 1$ , one can write

$$\mathbf{w}^H \mathbf{A}^H \mathbf{A} \mathbf{w} = \mathbf{w}_n^H \mathbf{a}_n \mathbf{a}_n^H \mathbf{w}. \quad (9.83)$$

Here,  $\mathbf{w}_n$  denotes the  $N_n \times 1$  vector of those entries of  $\mathbf{w}$  which contribute only to the  $n$ th tap of the end-to-end CIR  $h[\cdot]$  (or, equivalently, to the  $(n+1)$ th entry for the vector of the end-to-end CIR taps  $\mathbf{h}(\mathbf{w})$ ), the  $1 \times N_n$  row vector  $\mathbf{a}_n^H$  captures the nonzero<sup>4</sup> entries of the  $(n+1)$ th row of  $\mathbf{A}$  which correspond to the entries of  $\mathbf{w}_n$ , and  $N_n$  represents the number of the relays which contribute to the  $n$ th nonzero tap of the end-to-end CIR  $h[\cdot]$  (or, equivalently, to the  $(n+1)$ th nonzero entry for the vector of the end-to-end CIR taps  $\mathbf{h}(\mathbf{w})$ ). Based on Eq. (9.83), the optimization problem (9.82) can be written as

---

<sup>4</sup>If the  $(n+1)$ th row of matrix  $\mathbf{A}$  is zero, then the  $n$ th tap of the end-to-end CIR, i.e.,  $h[n]$  is zero. For this particular value of  $n$ ,  $\mathbf{a}_n$  and  $\mathbf{w}_n$  will be empty vectors and the inner maximization in the optimization problem (9.82) needs not be solved as the corresponding maximum balanced SNR for this tap  $n$  is 0.

$$\max_{n \in \mathcal{N}_c} \max_{\mathbf{w}_n} \frac{\left( P_T^{\max} / \sigma^2 - \mathbf{w}_n^H \mathbf{w}_n \right) \mathbf{w}_n^H \mathbf{a}_n \mathbf{a}_n^H \mathbf{w}}{2 \left[ \left( \mathbf{w}_n^H \mathbf{D}_1^{(n)} \mathbf{w}_n + 1 \right) \left( \mathbf{w}_n^H \mathbf{D}_2^{(n)} \mathbf{w}_n + 1 \right) \right]} \text{ subject to } \mathbf{w}_n^H \mathbf{w}_n \leq \frac{P_T^{\max}}{\sigma^2}, \quad (9.84)$$

Here, for  $q = 1, 2$ ,  $\mathbf{D}_q^{(n)}$  is an  $N_n \times N_n$  diagonal matrix whose diagonal entries are those diagonal entries of matrix  $\mathbf{D}_q$  that correspond to the relays which contribute to the  $n$ th tap of the end-to-end CIR. The inner maximization in the optimization problem (9.84) has the same structure as the maximization problem in the optimization problem (9.57). Hence, for each nonzero tap of the end-to-end CIR, an algorithm similar to the semi-closed-form technique presented in Section 9.4.1.2 under Result 9.16 can be used to obtain the corresponding maximum balanced SNR. Note that the nonzero rows of matrix  $\mathbf{A}$  determine the potentially nonzero taps, i.e., the potentially nonzero entries of the vector of the end-to-end channel taps  $\mathbf{h}(\mathbf{w}) = \mathbf{A}\mathbf{w}$ . If the  $(n+1)$ th row of matrix  $\mathbf{A}$  is zero, then the  $n$ th tap of the end-to-end CIR,  $h[\cdot]$  is zero as well. From Section 9.4.1.2, the max-min SNR optimal value of  $\mathbf{w}_n$  has a semi-closed-form solution given by

$$\mathbf{w}_n^o = \kappa_n \sqrt{2\nu_n} \left( 2\mu_n \mathbf{D}_1^{(n)} + 2\nu_n \mathbf{D}_2^{(n)} + \mathbf{I}_{N_n} \right)^{-1} \mathbf{a}_n. \quad (9.85)$$

Here,  $\nu_n \triangleq 0.5P_T^{\max}/\sigma^2 - \mu_n$ ,  $\kappa_n$  is defined as

$$\kappa_n \triangleq \left( \mathbf{a}_n^H \left( \mathbf{I}_{N_n} + 2\mu_n \mathbf{D}_1^{(n)} \right) \left( 2\mu_n \mathbf{D}_1^{(n)} + 2\nu_n \mathbf{D}_2^{(n)} + \mathbf{I}_{N_n} \right)^{-2} \mathbf{a}_n \right)^{-1/2}, \quad (9.86)$$

and  $\mu_n$  is the *provably unique* solution to the following equation:

$$\begin{aligned} & \left( \frac{P_T^{\max}}{\sigma^2} - 4\mu_n \right) \mathbf{a}_n^H \left( 2\mu_n \mathbf{D}_1^{(n)} + \left( \frac{P_T^{\max}}{\sigma^2} - 2\mu_n \right) \mathbf{D}_2^{(n)} + \mathbf{I}_{N_n} \right)^{-1} \mathbf{a}_n \\ & - \mu_n \left( \frac{P_T^{\max}}{\sigma^2} - 2\mu_n \right) \mathbf{a}_n^H \left( 2\mu_n \mathbf{D}_1^{(n)} + \left( \frac{P_T^{\max}}{\sigma^2} - 2\mu_n \right) \mathbf{D}_2^{(n)} + \mathbf{I}_{N_n} \right)^{-2} \left( 2\mathbf{D}_1^{(n)} - 2\mathbf{D}_2^{(n)} \right) \mathbf{a}_n = 0, \end{aligned} \quad (9.87)$$

which satisfies  $0 \leq \mu_n \leq 0.5P_T^{\max}/\sigma^2$ . To solve Eq. (9.87), a simple bisection method can be used to find the value of  $\mu_n$  in the interval  $[0, 0.5P_T^{\max}/\sigma^2]$  for which the left hand side (LHS) of Eq. (9.87) changes sign. In fact, the LHS of Eq. (9.87) is positive/negative for those values of  $\mu_n$  which are smaller/larger than the unique solution to Eq. (9.87).

When the  $n$ th tap of  $h[\cdot]$  is nonzero (i.e., when  $\mathbf{w}_n$  is chosen as in Eq. (9.85) and when the rest of the relays are switched off), the corresponding maximum balanced SNR, i.e., the value of the inner maximization in the optimization problem (9.84) is given by (see Eq. 9.61)

$$\text{SNR}_{\max}^{(n)} = \mu_n \left( P_T^{\max} / \sigma^2 - 2\mu_n \right) \mathbf{a}_n^H \left( 2\mu_n \mathbf{D}_1^{(n)} + 2\nu_n \mathbf{D}_2^{(n)} + \mathbf{I}_{N_n} \right)^{-1} \mathbf{a}_n. \quad (9.88)$$

Solving the outer maximization in Eq. (9.84) amounts to calculating the value of  $\text{SNR}_{\max}^{(n)}$  for all possible values of  $n \in \{0, 1, \dots, N_c - 1\}$ . The value of  $n$  which results in the largest value for  $\text{SNR}_{\max}^{(n)}$  is introduced as the only nonzero tap of the end-to-end CIR. That is, the optimal value of  $n$  is obtained as

$$n^o = \arg \max_{n \in \mathcal{N}_c} \text{SNR}_{\max}^{(n)}. \quad (9.89)$$

Once  $n^o$  is obtained, the optimal value  $\mathbf{w}_{n^o}^o$  is calculated from Eq. (9.85) by replacing  $n$  with  $n^o$ . Let  $\mathbf{w}^o$  represent the optimal value of the relay weight vector. If the  $r$ th relay contributes to tap  $n^o$  of the end-to-end CIR, then the  $r$ th entry of  $\mathbf{w}^o$  is equal to the entry of  $\mathbf{w}_{n^o}^o$  which corresponds to the  $r$ th relay. If the  $r$ th relay does not contribute to tap  $n^o$  of the end-to-end CIR, then the  $r$ th entry of  $\mathbf{w}^o$  is zero. With Eq. (9.78), the optimal value of  $P_{iq}$  can be calculated as

$$\begin{aligned} P_{iq}^o &= \frac{N(P_T^{\max} - \sigma^2 \mathbf{w}^H \mathbf{w})}{2(\mathbf{w}^H \mathbf{D}_q \mathbf{w} + 1) |\mathbf{f}_i^H \tilde{\mathbf{h}}(\mathbf{w})|^2 \mathbf{1}^T \mathbf{u}(\mathbf{w})} \Big|_{\mathbf{w}=\mathbf{w}^o} \\ &= \frac{\sigma^2 (P_T^{\max} - \sigma^2 \|\mathbf{w}^o\|^2)}{2(\mathbf{w}^{o,H} \mathbf{D}_q \mathbf{w}^o + 1)} = \frac{\sigma^2 (P_T^{\max} - \sigma^2 \|\mathbf{w}_{n^o}^o\|^2)}{2(\mathbf{w}_{n^o}^{o,H} \mathbf{D}_q^{(n)} \mathbf{w}_{n^o}^o + 1)}, \end{aligned} \quad (9.90)$$

where the fact that at the optimum  $|\mathbf{f}_i^H \tilde{\mathbf{h}}(\mathbf{w}^o)| = |\mathbf{f}_j^H \tilde{\mathbf{h}}(\mathbf{w}^o)|$  holds true, has been used to write,  $\mathbf{u}(\mathbf{w}^o) = \frac{1}{|\mathbf{f}_i^H \tilde{\mathbf{h}}(\mathbf{w}^o)|^2} \mathbf{1}$ . Interestingly, it follows from Eq. (9.90) that the subcarrier powers for each transceiver are all the same.

Based on the max-min SNR approach, the optimal joint power control and network beamforming for the multi-carrier equalization scheme is summarized in the following result.

### RESULT 9.23

For multi-carrier asynchronous two-way relay networks, the solution to the max-min SNR approach to network beamforming and power allocation is summarized below.

- Step 1. Set  $n = 0$ .
- Step 2. If no relay contributes to the  $n$ th tap of  $h[\cdot]$  (i.e., if the  $(n+1)$ th row of matrix  $\mathbf{A}$  is zero), let  $\text{SNR}_{\max}^{(n)} = 0$  and go to Step 6. Otherwise go to Step 3.
- Step 3. Let  $\mathbf{a}_n^H$  capture the nonzero entries of the  $(n+1)$ th row of  $\mathbf{A}$ .
- Step 4. Use a bisection algorithm to obtain the solution to Eq. (9.87) for  $\mu_n$  in the interval  $[0 \ 0.5P_T^{\max}/\sigma^2]$  and calculate  $\nu_n = 0.5P_T^{\max}/\sigma^2 - \mu_n$ .
- Step 5. Use Eq. (9.88) to calculate  $\text{SNR}_{\max}^{(n)}$ .
- Step 6. Let  $n = n + 1$ . If  $n \geq N_c$ , go to Step 7. Otherwise go to Step 2.
- Step 7. Find  $n^o$  such that  $\text{SNR}_{\max}^{(n^o)} \geq \text{SNR}_{\max}^{(n)}$ , for  $n = 0, 1, \dots, N_c - 1$ , that is  $n^o = \arg \max_{n \in \mathcal{N}_c} \text{SNR}_{\max}^{(n)}$ .
- Step 8. Use Eqs. (9.85)–(9.87) to calculate the optimal value of  $\mathbf{w}_{n^o}^o$ , where  $\nu_{n^o} = 0.5P_T^{\max}/\sigma^2 - \mu_{n^o}$ .
- Step 9. Let  $\mathbf{w}^o$  denote the optimal relay weight vector. If the  $r$ th relay is active, the  $r$ th entry of  $\mathbf{w}^o$  is equal to the element of  $\mathbf{w}_{n^o}^o$  which corresponds to the  $r$ th relay. If the  $r$ th relay is not active, then the  $r$ th entry of  $\mathbf{w}^o$  is zero.
- Step 10. Calculate the transceivers' subcarrier powers as in Eq. (9.90).

Assuming distributed beamforming at the relays and power control at the transceivers, the study in [55] characterizes the achievable SNR region and the corresponding rate region for such networks. Such a characterization is performed when each subcarrier is used to establish a bidirectional communication

between several outer transceivers. This study proves that if the rates over different subcarriers at each transceiver are assumed to be equal, the SNR region characterization leads to semi-closed-form solutions for the relay beamforming weights and transceivers' subcarrier powers and for the boundaries of the SNR region. Indeed, these solutions are identical to what is summarized under [Result 9.23](#).

#### 9.4.2.4 Sum-rate maximization approach

An alternative approach to design a network beamformer is the sum-rate maximization approach, subject to a total power constraint. In this approach, the design parameters are obtained as the solution to the following maximization:

$$\max_{\mathbf{p}_1, \mathbf{p}_2 \geq 0} \max_{\mathbf{w}} \sum_{i=1}^{N_f} \sum_{q=1}^2 \frac{1}{2} \log(1 + \text{SNR}_{iq}(P_{iq}, \mathbf{w})) \quad \text{subject to} \quad \frac{\mathbf{1}^T \mathbf{p}_1}{N_f} + \frac{\mathbf{1}^T \mathbf{p}_2}{N_f} + \sum_{r=1}^{N_r} Q_r \leq P_T^{\max}. \quad (9.91)$$

The following interesting result is proven in [56].

#### RESULT 9.24

In two-way multi-carrier asynchronous relay networks, under a total power constraint, the solution to the sum-rate maximization problem of joint network beamforming and power allocation is identical to the max-min SNR fair design approach to this problem as presented in [Result 9.23](#).

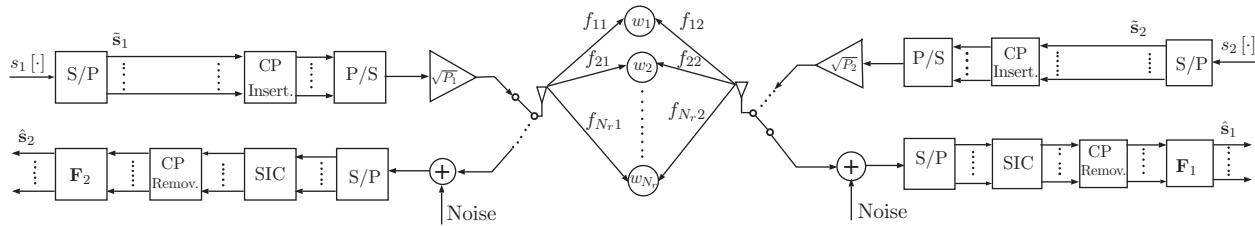
Earlier, [Result 9.20](#) established the equivalence of the max-min SNR design approach and the sum-rate maximization technique under a total power budget for synchronous two-way relay networks. [Result 9.24](#) establishes the same equivalency for multi-carrier asynchronous relay networks.

It is noteworthy that since the solution in [Result 9.24](#) leads to a single-tap end-to-end CIR, the *sum-rate-optimal (or max-min-optimal)* end-to-end channel is frequency-flat, and hence, the length of the CP prefix can be as small as zero.

#### 9.4.2.5 Single-carrier post-channel equalization

As an alternative approach to multi-carrier equalization, single-carrier equalization can be used at the two transceivers to compensate the end-to-end CIR. In such an approach, assuming a block equalization technique, the end-to-end channel is compensated either through post-channel equalization or through pre-channel equalization. In both schemes, cyclic prefix can be inserted to avoid inter-block-interference caused by time-dispersion of the end-to-end CIR.<sup>5</sup> A post-channel equalization scheme is shown in [Fig. 9.5](#): at the transmitter front-end of each transceiver, the stream of information symbols is organized into blocks. The blocks are

<sup>5</sup>As will be presented later, similar to the multi-carrier scheme presented in the previous subsection, the *MSE-optimal, power-optimal, and sum-rate-optimal* end-to-end CIR will have only one nonzero tap, and hence, the length of the cyclic prefix can be as small as zero.

**FIG. 9.5**

Block diagram of a single-carrier post-channel equalization-based two-way relay network [48].

then appended with the CP, and are then serially transmitted, after power adjustment, over the asynchronous two-way relay channel. At the receiver front-end of each transceiver, the received signals are turned into blocks of signals, are passed through the self-interference cancelation block, denoted as “SIC”, and are stripped off of the CP. Last but not least, the so-obtained blocks of received signals are linearly transformed (via multiplication by the corresponding post-channel equalizing matrix  $\mathbf{F}_1$  or  $\mathbf{F}_2$ ) with the aim to obtain linear estimates of the transmitted symbol blocks.

At the input of the post-channel equalizer of Transceiver  $q$ , the  $N_s \times 1$  vector of the received signals, denoted as  $\tilde{\mathbf{r}}_q$ , is expressed as

$$\tilde{\mathbf{r}}_q = \sqrt{P_{\bar{q}}} \tilde{\mathbf{H}}(\mathbf{w}) \tilde{\mathbf{s}}_{\bar{q}} + \text{noise}. \quad (9.92)$$

Here,  $\tilde{\mathbf{s}}_{\bar{q}}$  is the  $N_s \times 1$  vector of the symbols transmitted by Transceiver  $\bar{q}$ ,  $\tilde{\mathbf{H}}(\mathbf{w})$  is an  $N_s \times N_s$  circulant matrix whose  $(k, l)$ th entry is given as  $\check{h}[(k-l) \bmod N_s]$ , where  $\check{h}[n] = h[n]$ , if  $0 \leq n \leq N_c - 1$ , and  $\check{h}[n] = 0$ , if  $N_c \leq n \leq N_s - 1$ , and  $N_s$  is the number of symbols in each transmitted block. The noise term in Eq. (9.92) is the sum of two noise vectors: the vector of received noises at Transceiver  $q$  and a vector which depends on the relay noises. As shown in Fig. 9.5, this is how the latter noise vector is formed: at each relay, the received noise is multiplied by the relay beamforming weight. This weighted noise then goes through the channel between that relay and Transceiver  $\bar{q}$ , and thus, is delayed and attenuated accordingly. The delayed and attenuated versions of the weighted noises of different relays add up at Transceiver  $\bar{q}$ . This relay noise mixture received at Transceiver  $\bar{q}$  goes through the “S/P” block (i.e., it is turned into a vector), and then, passes through the CP removal block.

In the post-channel equalization approach, at Transceiver  $q$ , the received signals corresponding to one transmitted block of symbols (which contains  $N_s$  symbols) are linearly transformed, through multiplication of the vector of the received signal by an equalization matrix, thereby providing a linear estimate of the transmitted symbol vector. In this approach, the design parameters are the transceivers’ transmit powers  $P_1$  and  $P_2$ , the relay beamforming vector  $\mathbf{w}$ , and the post-channel equalization matrices, which are denoted as  $\mathbf{F}_1$  and  $\mathbf{F}_2$ . Three design approaches have been studied in the literature: (1) a total MSE minimization approach under a total power constraint, (2) a sum-rate maximization technique under a total power constraint, (3) a total power minimization approach under individual rate or MSE constraint. Below, we review the results obtained in these approaches.

#### 9.4.2.6 Total MSE minimization

In the total (or sum) MSE minimization approach, the design parameters are obtained as the solution to the following optimization problem:

$$\min_{\substack{P_1 \geq 0 \\ P_2 \geq 0}} \min_{\mathbf{w}} \min_{\mathbf{F}_1, \mathbf{F}_2} \sum_{q=1}^2 \text{MSE}_q(P_{\bar{q}}, \mathbf{F}_q, \mathbf{w}), \quad \text{subject to } P_T \leq P_T^{\max}, \quad (9.93)$$

where, as shown in [48], the total transmit power consumed in the entire network is given as  $P_T = P_1(1 + \mathbf{w}^H \mathbf{D}_1 \mathbf{w}) + P_2(1 + \mathbf{w}^H \mathbf{D}_2 \mathbf{w}) + \sigma^2 \mathbf{w}^H \mathbf{w}$ , while  $\text{MSE}_q(P_{\bar{q}}, \mathbf{F}_q, \mathbf{w})$  is

the MSE between the symbol vectors transmitted by Transceiver  $\bar{q}$  and their linear estimates, and is given as [48]

$$\begin{aligned} \text{MSE}_q(P_{\bar{q}}, \mathbf{F}_q, \mathbf{w}) &\triangleq \mathbb{E}\{\|\tilde{\mathbf{s}}_{\bar{q}} - \mathbf{F}_q \tilde{\mathbf{r}}_q\|^2\} \\ &= \text{tr}\left\{\mathbf{F}_q \mathbf{R}_q(\mathbf{w}) \mathbf{F}_q^H\right\} - \sqrt{P_{\bar{q}}} \text{tr}\left\{\mathbf{F}_q \tilde{\mathbf{H}}(\mathbf{w}) + \tilde{\mathbf{H}}^H(\mathbf{w}) \mathbf{F}_q^H\right\} + N_s. \end{aligned} \quad (9.94)$$

Here,  $\mathbf{R}_q(\mathbf{w})$  is the correlation matrix of the data vector  $\tilde{\mathbf{r}}_q$  received at Transceiver  $q$  and can be written, based on Eq. (9.92), as [48]

$$\mathbf{R}_q(\mathbf{w}) = P_{\bar{q}} \tilde{\mathbf{H}}(\mathbf{w}) \tilde{\mathbf{H}}^H(\mathbf{w}) + \sigma^2 (\mathbf{w}^H \mathbf{D}_q \mathbf{w} + 1) \mathbf{I}_{N_s}. \quad (9.95)$$

Note that the  $N_s \times N_s$  circulant matrix  $\tilde{\mathbf{H}}(\mathbf{w})$  can be decomposed as

$$\tilde{\mathbf{H}}(\mathbf{w}) = \mathbf{F}^H \mathbf{D}(\mathbf{w}) \mathbf{F}, \quad (9.96)$$

where

$$\mathbf{D}(\mathbf{w}) \triangleq \text{diag}\left\{H(e^{j0}), H(e^{j\frac{2\pi}{N_s}}), \dots, H(e^{j\frac{2\pi(N_s-1)}{N_s}})\right\} = \sqrt{N_s} \text{ diag}\left\{\mathfrak{f}_1^H \check{\mathbf{h}}(\mathbf{w}), \mathfrak{f}_2^H \check{\mathbf{h}}(\mathbf{w}), \dots, \mathfrak{f}_{N_s}^H \check{\mathbf{h}}(\mathbf{w})\right\}$$

is an  $N_s \times N_s$  diagonal matrix of the values of the frequency response of the end-to-end channel at integer multiples of  $1/N_s$ , whereas  $H(e^{j2\pi f}) \triangleq \sum_{n=0}^{N_s-1} h[n] e^{-jn2\pi f}$  is the frequency response of the end-to-end channel at the normalized frequency  $f$ , matrix  $\mathbf{F}$  is here redefined as the  $N_s \times N_s$  DFT matrix whose  $(k, k')$ th element is defined as  $F(k, k') = N_s^{-\frac{1}{2}} e^{-j2\pi(k-1)(k'-1)/N_s}$ , for  $k, k' \in \mathcal{N}_s \triangleq \{1, \dots, N_s\}$ , vector  $\mathfrak{f}_k$  is redefined as

$$\mathfrak{f}_k \triangleq \frac{1}{\sqrt{N_s}} \left[ 1 \ e^{\left(j\frac{2\pi(k-1)}{N_s}\right)} \ e^{\left(j\frac{4\pi(k-1)}{N_s}\right)} \ \dots \ e^{\left(j\frac{2(N_s-1)(k-1)\pi}{N_s}\right)} \right]^T,$$

which is the  $k$ th row of  $\mathbf{F}^H$ , for  $k \in \mathcal{N}_s$ , and  $\check{\mathbf{h}}(\mathbf{w}) \triangleq [\mathbf{h}^T(\mathbf{w}) \ \mathbf{0}_{1 \times (N_s-N_c)}]^T$  is the zero-padded version of the vector of the channel taps  $\mathbf{h}(\mathbf{w})$ . For any given values of  $P_1, P_2$ , and  $\mathbf{w}$ , differentiating the cost function in Eq. (9.93) with respect to  $\mathbf{F}_q$  and equating the derivative to zero yield the optimal value of  $\mathbf{F}_q$  as

$$\mathbf{F}_q^o(P_{\bar{q}}, \mathbf{w}) = \sqrt{P_{\bar{q}}} \tilde{\mathbf{H}}^H(\mathbf{w}) \mathbf{R}_q^{-1}(\mathbf{w}). \quad (9.97)$$

Based on Eqs. (9.96) and (9.97), the total MSE minimization problem (9.93) can be simplified as [48]

$$\begin{aligned} \min_{\substack{P_1 \geq 0 \\ P_2 \geq 0}} \min_{\mathbf{w}} \quad & \sum_{q=1}^2 \sum_{k=1}^{N_s} \frac{1}{\frac{P_{\bar{q}} N_s |\mathfrak{f}_k^H \check{\mathbf{h}}(\mathbf{w})|^2}{\sigma^2 (\mathbf{w}^H \mathbf{D}_q \mathbf{w} + 1)} + 1} \\ \text{subject to} \quad & \sum_{q=1}^2 P_q (1 + \mathbf{w}^H \mathbf{D}_q \mathbf{w}) + \sigma^2 \mathbf{w}^H \mathbf{w} \leq P_T^{\max}. \end{aligned} \quad (9.98)$$

It is proven in [48] that at the optimum of optimization problem (9.98), the following result is true.

**RESULT 9.25**

The solution to the optimization problem (9.98) is such that only one of the entries of the vector of the taps of the end-to-end CIR  $\mathbf{h}(\mathbf{w}) = \mathbf{Aw}$  is nonzero.

**Result 9.25** states that at the optimum of the optimization problem (9.93), only one of the taps of the end-to-end CIR is nonzero. Hence, in order to achieve the minimum total MSE, all relays which contribute to the zero taps of end-to-end CIR should be assigned a zero weight, i.e., those relays will not participate in the relaying, and only the relays which contribute to the only nonzero tap of the end-to-end CIR  $h[\cdot]$  will participate in the relaying. Recall that  $\mathcal{U}_n$  represents the set of those values of  $\mathbf{w}$  which have zero entries for those relays which do not contribute to the  $n$ th of the end-to-end CIR and which could have nonzero zero entries for those relays which do contribute to the  $n$ th of the end-to-end CIR. **Result 9.25** implies that at the optimum of the optimization problem (9.93),  $\mathbf{w} \in \bigcup_{n=0}^{N_c-1} \mathcal{U}_n$  holds true, and thus,  $|\mathbf{f}_k^H \check{\mathbf{h}}(\mathbf{w})|^2 = |\mathbf{f}_{k'}^H \check{\mathbf{h}}(\mathbf{w})|^2$  holds true for  $k \neq k'$ . Hence, for  $\mathbf{w} \in \bigcup_{n=0}^{N_c-1} \mathcal{U}_n$ , the cost function in the optimization problem (9.98) can be written as

$$\sum_{q=1}^2 \sum_{k=1}^{N_s} \frac{1}{P_q N_s |\mathbf{f}_k^H \check{\mathbf{h}}(\mathbf{w})|^2 + 1} = \sum_{q=1}^2 \frac{N_s}{P_q N_s |\mathbf{f}_k^H \check{\mathbf{h}}(\mathbf{w})|^2 + 1} = \sum_{q=1}^2 \frac{N_s}{P_q \mathbf{w}^H \mathbf{A}^H \mathbf{A} \mathbf{w}} + 1, \quad (9.99)$$

where in the first equality,  $k$  can be any integer in the set  $\{1, 2, \dots, N_s\}$  and in the second equality, Parseval's identity is used. That is, for  $\mathbf{w} \in \bigcup_{n=0}^{N_c-1} \mathcal{U}_n$ , one can write

$$N_s |\mathbf{f}_k^H \check{\mathbf{h}}(\mathbf{w})|^2 = \sum_{k'=1}^{N_s} |\mathbf{f}_{k'}^H \check{\mathbf{h}}(\mathbf{w})|^2 = \check{\mathbf{h}}^H(\mathbf{w}) \check{\mathbf{h}}(\mathbf{w}) = \mathbf{h}^H(\mathbf{w}) \mathbf{h}(\mathbf{w}) = \mathbf{w}^H \mathbf{A}^H \mathbf{A} \mathbf{w}. \quad (9.100)$$

To further simplify the cost function in the optimization problem (9.99), it can be shown that at the optimum, the following equality holds:

$$\frac{P_1 \mathbf{w}^H \mathbf{A}^H \mathbf{A} \mathbf{w}}{\sigma^2 (\mathbf{w}^H \mathbf{D}_2 \mathbf{w} + 1)} = \frac{P_2 \mathbf{w}^H \mathbf{A}^H \mathbf{A} \mathbf{w}}{\sigma^2 (\mathbf{w}^H \mathbf{D}_1 \mathbf{w} + 1)}, \quad (9.101)$$

and hence, using the optimization problem (9.99) and (9.101) along with the fact that the optimal  $\mathbf{w}$  belongs to  $\mathcal{W} \triangleq \bigcup_{n=0}^{N_c-1} \mathcal{U}_n$ , one can rewrite the optimization problem (9.97), equivalently, as

$$\max_{P_1 \geq 0} \max_{\mathbf{w} \in \mathcal{W}} \frac{P_1 \mathbf{w}^H \mathbf{A}^H \mathbf{A} \mathbf{w}}{\sigma^2 (\mathbf{w}^H \mathbf{D}_2 \mathbf{w} + 1)} \text{ subject to } 2P_1 (1 + \mathbf{w}^H \mathbf{D}_1 \mathbf{w}) + \sigma^2 \mathbf{w}^H \mathbf{w} = P_T^{\max}, \quad (9.102)$$

or, equivalently as

$$\max_{n \in \mathcal{N}_c} \max_{\mathbf{w} \in \mathcal{U}_n} \frac{(P_T^{\max} - \sigma^2 \mathbf{w}^H \mathbf{w}) \mathbf{w}^H \mathbf{A}^H \mathbf{A} \mathbf{w}}{\sigma^2 (\mathbf{w}^H \mathbf{D}_1 \mathbf{w} + 1) (\mathbf{w}^H \mathbf{D}_2 \mathbf{w} + 1)}, \text{ subject to } \mathbf{w}^H \mathbf{w} \leq \frac{P_T^{\max}}{\sigma^2}. \quad (9.103)$$

The optimization problem (9.103) is identical to the optimization problem (9.82), leading to the following result:

### RESULT 9.26

The solution to the total MSE minimization problem (9.93), in terms of  $\mathbf{w}$ , can be obtained using Steps 1–9 of the algorithm presented under [Result 9.23](#). Based on Eq. (9.101) and the constraint in the optimization problem (9.102), the optimal value of the transceivers' transmit powers are obtained as

$$P_q^o = \frac{\sigma^2 (P_T^{\max} - \sigma^2 \| \mathbf{w}^o \|^2)}{2(\mathbf{w}^{o,H} \mathbf{D}_q \mathbf{w}^o + 1)} = \frac{\sigma^2 (P_T^{\max} - \sigma^2 \| \mathbf{w}_{n^o}^o \|^2)}{2(\mathbf{w}_{n^o}^{o,H} \mathbf{D}_q^{(n)} \mathbf{w}_{n^o}^o + 1)}, \text{ for } q = 1, 2. \quad (9.104)$$

The optimal values of the post-channel equalizers can be obtained from Eq. (9.97) by replacing  $\mathbf{w}$  with  $\mathbf{w}^o$ .

In other words, the MSE-optimal value of the beamforming weight vector in the single-carrier scheme is identical to the max-min-SNR-optimal value of the beamforming weight vector in the multi-carrier scheme. It is worth mentioning that for the single-carrier scheme, the SNR expression in Eq. (9.88) is the maximum SNR that can be achieved in the single-carrier scheme for every transmitted symbol when only the  $n$ th tap of the end-to-end CIR is nonzero.

There are two other MSE-based approaches to design joint network beamforming and power allocation for single-carrier asynchronous two-way relay networks, namely the min-max MSE approach and the MSE balancing approach, both under a total power budget constraint. The min-max MSE approach solves the following optimization problem:

$$\min_{\substack{P_2 \geq 0 \\ P_1 \geq 0}} \min_{\mathbf{w}} \max_{q \in \{1, 2\}} \min_{\mathbf{F}_q} \text{MSE}_q(P_{\bar{q}}, \mathbf{F}_q, \mathbf{w}), \text{ subject to } P_T \leq P_T^{\max}. \quad (9.105)$$

The MSE balancing method solves the following optimization problem:

$$\begin{aligned} & \min_{\substack{P_2 \geq 0 \\ P_2 \geq 0}} \min_{\mathbf{w}} \min_{\mathbf{F}_1} \text{MSE}_1(P_2, \mathbf{F}_1, \mathbf{w}), \\ & \text{subject to } P_T \leq P_T^{\max} \text{ and } \min_{\mathbf{F}_1} \text{MSE}_1(P_2, \mathbf{F}_1, \mathbf{w}) = \min_{\mathbf{F}_2} \text{MSE}_2(P_1, \mathbf{F}_2, \mathbf{w}). \end{aligned} \quad (9.106)$$

The following result is proven in [48]:

### RESULT 9.27

The solution to the min-max MSE problem (9.105) and the solution to the MSE balancing problem (9.106) are identical to the solution to the total MSE minimization problem (9.93). Hence, these solutions can be obtained by using the algorithm summarized under [Result 9.26](#).

#### 9.4.2.7 Sum-rate maximization

In this subsection, the focus is to present a method which optimally obtains the relay beamforming weight vector  $\mathbf{w}$  and the transceivers' transmit powers  $P_1$  and  $P_2$  such that the sum-rate is maximized under a total transmit power budget. Note that in this design approach, no equalization method is assumed. Indeed, this approach aims to

find the highest sum-rate which can be achieved with any equalization scheme. It is shown in [57] that the data model for the single-carrier scheme can be viewed as a multiple-input multiple-output (MIMO) channel. Based on this viewpoint, one can use Eq. (9.92) to express the data rate, which can be achieved at Transceiver  $q$ , as

$$R_q(P_{\bar{q}}, \mathbf{w}) = \frac{1}{2} \log \det \left\{ \mathbf{I}_{N_s} + P_{\bar{q}} \mathbf{R}_q^{-1/2}(\mathbf{w}) \tilde{\mathbf{H}}(\mathbf{w}) \tilde{\mathbf{H}}^H(\mathbf{w}) \mathbf{R}_q^{-1/2}(\mathbf{w}) \right\}, \text{ for } q = 1, 2, \quad (9.107)$$

where  $\mathbf{R}_q(\mathbf{w})$  and  $\tilde{\mathbf{H}}(\mathbf{w})$  are given as in Eqs. (9.95) and (9.96), respectively. To obtain the optimal values of the transceivers' transmit powers and the relay beamforming weight vector, based on Eq. (9.107), the problem of maximizing the sum-rate under a total power constraint can be formulated as<sup>6</sup>

$$\begin{aligned} & \max_{P_1, P_2, \mathbf{w}} \sum_{q=1}^2 \log \det \left\{ \mathbf{I}_{N_s} + P_{\bar{q}} \mathbf{R}_q^{-1/2}(\mathbf{w}) \tilde{\mathbf{H}}(\mathbf{w}) \tilde{\mathbf{H}}^H(\mathbf{w}) \mathbf{R}_q^{-1/2}(\mathbf{w}) \right\} \\ & \text{subject to } P_T \leq P_T^{\max}, \quad P_1 \geq 0, P_2 \geq 0. \end{aligned} \quad (9.108)$$

Indeed, each of the two terms in the objective function in the optimization problem (9.108) is the achievable rate at one of the transceivers. With Eq. (9.96), the optimization problem (9.108) can be written as [57]

$$\begin{aligned} & \max_{\mathbf{w}} \max_{P_1, P_2} \sum_{q=1}^2 \sum_{k=1}^{N_s} \log \left( 1 + \frac{P_{\bar{q}}}{\sigma^2(1+\mathbf{w}^H \mathbf{D}_q \mathbf{w})} N_s |\mathbf{f}_k^H \tilde{\mathbf{h}}(\mathbf{w})|^2 \right) \\ & \text{subject to } \sum_{q=1}^2 P_q (1 + \mathbf{w}^H \mathbf{D}_q \mathbf{w}) + \sigma^2 \mathbf{w}^H \mathbf{w} \leq P_T^{\max}, \quad P_1 \geq 0, P_2 \geq 0. \end{aligned} \quad (9.109)$$

To solve the optimization problem (9.109), one can solve the maximization over  $P_1$  and  $P_2$  while assuming that  $\mathbf{w}$  is fixed, and later obtain the optimal value of  $\mathbf{w}$ . Based on the method of Lagrange multipliers,<sup>7</sup> the solution to the inner maximization in the optimization problem (9.109) can be obtained as

$$P_q = \frac{P_T^{\max} - \sigma^2 \mathbf{w}^H \mathbf{w}}{2(1 + \mathbf{w}^H \mathbf{D}_q \mathbf{w})} \geq 0, \text{ for } q = 1, 2. \quad (9.110)$$

With Eq. (9.110), the optimization problem (9.109) is then written as

$$\max_{\mathbf{w}} \log \prod_{k=1}^{N_s} \left( 1 + \frac{(P_T^{\max} - \sigma^2 \mathbf{w}^H \mathbf{w}) N_s |\mathbf{f}_k^H \tilde{\mathbf{h}}(\mathbf{w})|^2}{2\sigma^2(1 + \mathbf{w}^H \mathbf{D}_1 \mathbf{w})(1 + \mathbf{w}^H \mathbf{D}_2 \mathbf{w})} \right)^2, \quad \text{subject to } \mathbf{w}^H \mathbf{w} \leq \frac{P_T^{\max}}{\sigma^2}. \quad (9.111)$$

The following result facilitates the simplification of the optimization problem (9.111).

---

<sup>6</sup>Here, the factor 1/2 in the rate expression in Eq. (9.107) is ignored.

<sup>7</sup>Indeed, one can easily show that at the optimum, the total power constraint is satisfied with equality, and hence, for fixed  $\mathbf{w}$ , the method of Lagrange multipliers is applicable to the inner maximization in the optimization problem (9.109).

**RESULT 9.28**

The solution to the optimization problem (9.111) is such that only one of the entries of the vector of the taps of the end-to-end CIR  $\mathbf{h}(\mathbf{w}) = \mathbf{A}\mathbf{w}$  is nonzero.

**Result 9.28** states that at the optimum of the sum-rate maximization problem (9.108), only one of the taps of the end-to-end CIR is nonzero. Hence, in order to achieve maximum sum-rate, all relays which contribute to the zero taps of end-to-end CIR should be assigned a zero weight, i.e., those relays will not participate in relaying, and only the relays which contribute to the only nonzero tap of the end-to-end CIR  $h[\cdot]$  will participate in relaying. Recall that  $\mathcal{U}_n$  represents the set of those values of  $\mathbf{w}$  which have zero entries for those relays which do not contribute to the  $n$ th of the end-to-end CIR and which could have nonzero entries for those relays which do contribute to the  $n$ th of the end-to-end CIR. **Result 9.28** implies that at the optimum of the optimization problem (9.108),  $\mathbf{w} \in \bigcup_{n=0}^{N_c-1} \mathcal{U}_n$  holds true, and thus, based on Eq. (9.100), the optimization problem (9.111) can be equivalently written as

$$\max_{\mathbf{w} \in \mathcal{N}_c} \max_{\mathbf{w} \in \mathcal{U}_n} \frac{(P_T^{\max} - \sigma^2 \mathbf{w}^H \mathbf{w}) \mathbf{w}^H \mathbf{A}^H \mathbf{A} \mathbf{w}}{2\sigma^2 (1 + \mathbf{w}^H \mathbf{D}_1 \mathbf{w})(1 + \mathbf{w}^H \mathbf{D}_2 \mathbf{w})}, \text{ subject to } \mathbf{w}^H \mathbf{w} \leq \frac{P_T^{\max}}{\sigma^2}. \quad (9.112)$$

The optimization problem (9.112) is identical to the optimization problem (9.82) and thus the following result holds.

**RESULT 9.29**

For single-carrier asynchronous two-way relay networks with post-channel equalization and under a total power budget constraint, the total-MSE optimal solution is sum-rate optimal. As such, the solution to the sum-rate optimization problem (9.108) is given in **Result 9.26**.

**9.4.2.8 Total power minimization**

A third design approach to find the relay beamforming weight vector  $\mathbf{w}$  and the transceivers' transmit powers  $P_1$  and  $P_2$  in a single-carrier asynchronous two-way relay network is to minimize the total power consumed in the entire network subject to two constraints on the transceivers' data rates. Mathematically, in this approach, the following optimization problem is solved:

$$\min_{\mathbf{w}, P_1, P_2} P_T, \text{ subject to } R_1(P_2, \mathbf{w}) \geq r_1, R_2(P_1, \mathbf{w}) \geq r_2, P_1 \geq 0, P_2 \geq 0, \quad (9.113)$$

where  $R_q(P_q, \mathbf{w})$ , given in Eq. (9.107), is the data rate achieved at Transceiver  $q$ , for  $q = 1, 2$ , and  $r_q$  is the corresponding minimum required data rates at this transceiver. Based on Eqs. (9.95), (9.96) and (9.107), the optimization problem (9.113) can be simplified as

$$\begin{aligned} & \min_{\mathbf{w}, P_1, P_2} \sum_{q=1}^2 P_q (\mathbf{w}^H \mathbf{D}_q \mathbf{w} + 1) + \sigma^2 \mathbf{w}^H \mathbf{w}, \\ & \text{subject to } \log \left( \prod_{k=1}^{N_s} \left( 1 + \frac{P_{\bar{q}} N_s |\mathbf{f}_k^H \check{\mathbf{h}}(\mathbf{w})|^2}{\sigma^2 (1 + \mathbf{w}^H \mathbf{D}_q \mathbf{w})} \right) \right) = 2r_q, \text{ for } q = 1, 2, \quad P_1 \geq 0, \quad P_2 \geq 0. \end{aligned} \quad (9.114)$$

The following key result is proved in [58].

### RESULT 9.30

The solution to the optimization problem (9.114) is such that only one of the entries of the vector of the taps of the end-to-end CIR  $\mathbf{h}(\mathbf{w}) = \mathbf{Aw}$  is nonzero.

**Result 9.30** states that at the optimum only one of the taps of the end-to-end CIR is nonzero. Hence, in order to minimize the total power, all relays which contribute to the zero taps of end-to-end CIR should be assigned a zero weight, i.e., those relays will not participate in relaying, and only the relays which contribute to the only non-zero tap of the end-to-end CIR  $h[\cdot]$  will participate in relaying. Recall that  $\mathcal{U}_n$  represents the set of those values of  $\mathbf{w}$  which have zero entries for those relays which do not contribute to the  $n$ th of the end-to-end CIR and which could have nonzero entries for those relays which do contribute to the  $n$ th of the end-to-end CIR. **Result 9.30** implies that at the optimum of the optimization problem (9.113),  $\mathbf{w} \in \bigcup_{n=0}^{N_c-1} \mathcal{U}_n$  holds true, and thus,  $|\mathbf{f}_k^H \check{\mathbf{h}}(\mathbf{w})|^2 = |\mathbf{f}_{k'}^H \check{\mathbf{h}}(\mathbf{w})|^2$  holds true for  $k \neq k'$ . Using **Result 9.30**, one can easily use the constraints in the optimization problem (9.114) to prove that at the optimum, the following relationship between  $P_1$  and  $\mathbf{w}$  holds:

$$P_{\bar{q}} = \frac{\sigma^2 (2^{2r_q/N_s} - 1)(1 + \mathbf{w}^H \mathbf{D}_q \mathbf{w})}{N_s |\mathbf{f}_k^H \check{\mathbf{h}}(\mathbf{w})|^2} = \frac{\sigma^2 (2^{2r_q/N_s} - 1)(1 + \mathbf{w}^H \mathbf{D}_q \mathbf{w})}{\mathbf{w}^H \mathbf{A}^H \mathbf{A} \mathbf{w}}, \quad (9.115)$$

where the second equality follows from Eq. (9.100). In light of Eq. (9.115) along with the fact that the optimal  $\mathbf{w}$  belongs to the set  $\bigcup_{n=0}^{N_c-1} \mathcal{U}_n$ , the optimization problem (9.114) can be written as

$$\min_{n \in \mathcal{N}_c} \min_{\mathbf{w} \in \mathcal{U}_n} \sigma^2 \left( \frac{(2^{2r_1/N_s} + 2^{2r_2/N_s} - 2)(1 + \mathbf{w}^H \mathbf{D}_1 \mathbf{w})(1 + \mathbf{w}^H \mathbf{D}_2 \mathbf{w})}{\mathbf{w}^H \mathbf{A}^H \mathbf{A} \mathbf{w}} + \mathbf{w}^H \mathbf{w} \right). \quad (9.116)$$

Based on the fact that for  $\mathbf{w} \in \mathcal{U}_n$  along with Eq. (9.83) (which establishes that  $\mathbf{w}^H \mathbf{A}^H \mathbf{A} \mathbf{w} = \mathbf{w}_n^H \mathbf{a}_n \mathbf{a}_n^H \mathbf{w}_n$  holds true), the optimization problem (9.116) can be equivalently written as

$$\min_{n \in \mathcal{N}_c} \min_{\mathbf{w}_n} \sigma^2 \left( \frac{(2^{2r_1/N_s} + 2^{2r_2/N_s} - 2) \left( 1 + \mathbf{w}_n^H \mathbf{D}_1^{(n)} \mathbf{w}_n \right) \left( 1 + \mathbf{w}_n^H \mathbf{D}_2^{(n)} \mathbf{w}_n \right)}{\mathbf{w}_n^H \mathbf{a}_n \mathbf{a}_n^H \mathbf{w}_n} + \mathbf{w}_n^H \mathbf{w}_n \right), \quad (9.117)$$

where, as defined earlier,  $\mathbf{w}_n$  denotes the  $N_n \times 1$  vector of those entries of  $\mathbf{w}$  which contribute only to the  $n$ th tap of the end-to-end CIR  $h[\cdot]$  (or equivalently, to the  $(n+1)$ th entry for the vector of the end-to-end CIR taps  $\mathbf{h}(\mathbf{w})$ ). Interestingly, the inner

minimization in the optimization problem (9.117) is similar to the one in the optimization problem (9.49). Hence for any given value of  $n$ , a semi-closed-form algorithm similar to the one presented under [Result 9.11](#) can be used to obtain the optimal value of  $\mathbf{w}_n$ . The outer minimization in the optimization problem (9.117) aims to find the index of the only nonzero tap of the end-to-end CIR which results in the smallest value of the total power consumed in the entire network.

### RESULT 9.31

For single-carrier asynchronous two-way relay networks, the solution to the total power minimization approach to network beamforming and power allocation is summarized below.

- Step 1. Set  $n = 0$ .
- Step 2. If no relay contributes to the  $n$ th tap of  $h[\cdot]$  (i.e., if the  $(n + 1)$ th row of the matrix  $\mathbf{A}$  is zero), go to [Step 6](#).
- Step 3. Let  $\mathbf{a}_n^H$  be a  $1 \times N_n$  vector whose  $r$ th entry is equal to the  $r$ th nonzero entry of the  $(n + 1)$ th row of matrix  $\mathbf{A}$ , where  $N_n$  is the number of the nonzero entries of the  $(n + 1)$ th row of matrix  $\mathbf{A}$ . Define  $f_{rq}^{(n)}$  as the channel coefficient between Transceiver  $q$  and the  $r$ th relay which contributes to the  $n$ th tap of the end-to-end CIR. Let  $\mathbf{f}_1^{(n)}$  be the channel vector between Transceiver 1 and the relays which contribute to the  $n$ th tap and define  $\beta_1 \triangleq \left(2^{\frac{2r_1}{N_s}} - 1\right)$  and  $\beta_2 \triangleq \left(2^{\frac{2r_2}{N_s}} - 1\right)$ . Also, define  $\hbar_n(z)$  as

$$\hbar_n(z) \triangleq 1 - (\beta_1 + \beta_2) \frac{1/z^2 - \lambda_n \mathbf{a}_n^H \left( (\beta_1 + \beta_2) \mathbf{D}_2^{(n)} - \lambda_n (z \mathbf{D}_1^{(n)} + \mathbf{I}_{N_n}) \right)^{-2} \mathbf{D}_1^{(n)} \mathbf{a}_n}{\lambda_n^2 \mathbf{a}_n^H ((\beta_1 + \beta_2) \mathbf{D}_2^{(n)} + \lambda_n (z \mathbf{D}_1^{(n)} + \mathbf{I}_{N_n}))^{-2} (z \mathbf{D}_1^{(n)} + \mathbf{I}_{N_n}) \mathbf{a}_n},$$

for  $n = 0, 1, \dots, N_c - 1$ ,

where for any given value of  $z$ ,  $\lambda_n$  is the unique positive root of the following nonlinear equation:

$$\sum_{r=1}^{N_n} \frac{z |f_{r1}^{(n)} f_{r2}^{(n)}|^2}{(\beta_1 + \beta_2) |f_{r2}^{(n)}|^2 + \lambda_n (z |f_{r1}^{(n)}|^2 + 1)} = 1, \text{ for } n = 0, 1, \dots, N_c - 1.$$

- Step 4. Define  $p_l \triangleq \frac{\beta_1 + \beta_2}{\|\mathbf{f}_1^{(n)}\|^2}$ , choose  $p_u$  to be a sufficiently large number, and let  $\epsilon$  be an arbitrarily small real-valued scalar.
- Step 5. Set  $k = 1$  and choose  $z_n^{(k)} = (p_l + p_u)/2$ .
- Step 6. If  $\hbar_n(z_n^{(k)}) > 0$ , set  $p_u = z_n^{(k)}$ . If  $\hbar_n(z_n^{(k)}) < 0$ , set  $p_l = z_n^{(k)}$ . Calculate  $z_n^{(k+1)} = (p_l + p_u)/2$ .
- Step 7. If  $|z_n^{(k+1)} - z_n^{(k)}| > \epsilon$ , then  $k = k + 1$  and go to [Step 6](#). Otherwise set  $z_n = z_n^{(k+1)}$  and calculate  $\kappa_n$  using

$$\kappa_n = \frac{1}{\sqrt{z_n \mathbf{a}_n^H (z_n \mathbf{D}_1^{(n)} + \mathbf{I}_{N_n}) \left( (\beta_1 + \beta_2) \mathbf{D}_2^{(n)} + \lambda_n (z_n \mathbf{D}_1^{(n)} + \mathbf{I}_{N_n}) \right)^{-2} \mathbf{a}_n}}.$$

- Step 8. Obtain the optimal value of the weight vector  $\mathbf{w}_n$ , denoted as  $\mathbf{w}_n^o$ , as

$$\mathbf{w}_n^o = \kappa_n \sqrt{\frac{z_n (\beta_1 + \beta_2)}{\lambda_n}} \left( (\beta_1 + \beta_2) \mathbf{D}_2^{(n)} + \lambda_n (z_n \mathbf{D}_1^{(n)} + \mathbf{I}_{N_n}) \right)^{-1} \mathbf{a}_n.$$

**RESULT 9.31—CONT'D**

Step 9. Calculate the cost function  $f_n(\mathbf{w}_n^o)$  as

$$f_n(\mathbf{w}_n^o) = \sigma^2 \left( \frac{(\beta_1 + \beta_2) (\mathbf{w}_n^{o,H} \mathbf{D}_1^{(n)} \mathbf{w}_n^o + 1) (\mathbf{w}_n^{o,H} \mathbf{D}_2^{(n)} \mathbf{w}_n^o + 1)}{\mathbf{w}_n^{o,H} \mathbf{a}_n \mathbf{a}_n^H \mathbf{w}_n^o} + \mathbf{w}_n^{o,H} \mathbf{w}_n^o \right).$$

Step 10. Set  $n = n + 1$ . If  $n \geq N_n$  go to the next step, otherwise go to [Step 2](#).

Step 11. Find the optimal value of  $n$ , denoted as  $n^o$ , which yields the smallest value of  $f_n(\mathbf{w}_n^o)$ , i.e.,

$$n^o = \arg \min_{n \in \mathcal{N}_c} f_n(\mathbf{w}_n^o).$$

Step 12. Let  $\mathbf{w}^o$  denote the optimal relay weight vector. If the  $r$ th relay contributes to tap  $n^o$  of the end-to-end CIR, then the  $r$ th entry of  $\mathbf{w}^o$  is equal to the element of  $\mathbf{w}_{n^o}^o$  which corresponds to the  $r$ th relay and if the  $r$ th relay does not contribute to tap  $n^o$  of the end-to-end CIR, then the  $r$ th entry of  $\mathbf{w}^o$  is zero.

Step 13. Calculate the transceiver transmit powers as

$$P_1^o = \frac{\beta_2 \left( 1 + \mathbf{w}_{n^o}^{o,H} \mathbf{D}_2^{(n^o)} \mathbf{w}_{n^o}^o \right)}{\mathbf{w}_{n^o}^{o,H} \mathbf{a}_{n^o} \mathbf{a}_{n^o}^H \mathbf{w}_{n^o}^o}, \quad P_2^o = \frac{\beta_1 \left( 1 + \mathbf{w}_{n^o}^{o,H} \mathbf{D}_1^{(n^o)} \mathbf{w}_{n^o}^o \right)}{\mathbf{w}_{n^o}^{o,H} \mathbf{a}_{n^o} \mathbf{a}_{n^o}^H \mathbf{w}_{n^o}^o}.$$

**9.4.2.9 Single-carrier pre-channel equalization**

Assuming pre-channel equalization at the transmitter front-ends of the transceiver (i.e., precoding) *without any processing at the receiver front-ends*, the study in [49] aims to minimize the total MSE between the received signals and the transmitted symbols subject to a total power constraint. In this total MSE minimization approach, the design parameters are the transmit powers of the two transceivers, the precoding matrices used at the transmitter front-ends of the two transceivers, and the relay beamforming weight vector.

The following result is rigorously proven in [49]:

**RESULT 9.32**

Assuming pre-channel equalization at the transmitter front-ends of the transceiver (i.e., precoding) *without any processing at the receive fronts-ends* and minimizing the total MSE (over the transmit powers of the two transceivers, the precoding matrices used at the transmitter front-ends of the two transceivers, and the relay beamforming weight vectors) under a total power budget, results in a relay selection scheme, where only the relays contributing to one tap of the end-to-end CIR, are turned on and the remaining relays are switched off.

The investigation in [49] presents an efficient method to obtain the index of the nonzero tap of the end-to-end CIR and the optimal values of the design parameters. The simulation results of [49] show that the pre-channel equalization technique

performs close to the post-channel equalization approach (presented in the previous subsection), when the total available power is relatively low compared to the noise power at the transceivers, while offering receiver simplicity. If however in the pre-channel equalization approach, a simple amplification factor<sup>8</sup> is allowed at each receiver front-end of the two transceivers, one can then prove that the pre-channel equalization and the post-channel equalization scheme lead to the same solution.

#### 9.4.2.10 Joint pre-channel and post-channel equalization

In Section 9.4.2.2, the multi-carrier (OFDM-based) equalization scheme is considered and two approaches to the joint subcarrier power loading and network beamforming were presented: (1) the max-min SNR design-based approach under a total power constraint, and (2) the sum-rate maximization under a total power constraint. These approaches are shown to lead to the same solution. The OFDM-based equalization scheme belongs to the class of joint pre-channel and post-channel equalization schemes, where the precoding matrices used at the transmitter front-ends of the two transceivers and the channel equalizing matrices (linear estimators) used at the receiver front-ends of the two transceivers are inverse DFT and DFT matrices, respectively. A natural question that comes to mind is what are the MSE-optimal values of the precoding matrices used at the transmitter front-ends of the two transceivers and the optimal values of the channel equalizing matrices used at the receiver front-ends of the two transceivers. More specifically, given a certain optimality criterion, such as total MSE or sum-rate, the question worth answering is what is the optimal solution to jointly optimal precoding, linear symbol precoding and estimation, power loading and network beamforming.

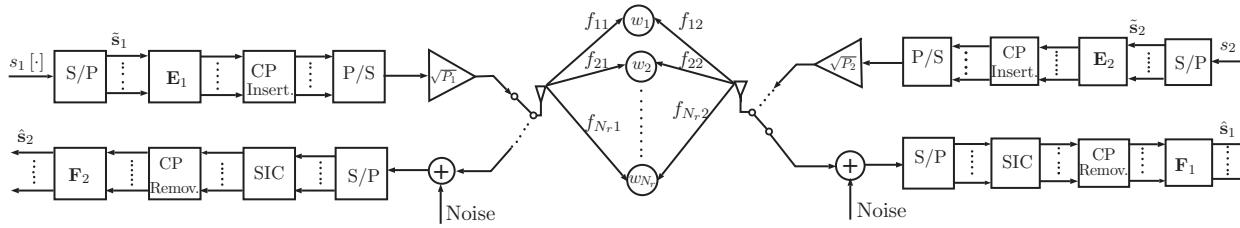
The corresponding communication scheme is shown in Fig. 9.6: at the transmitter front-end of each transceiver, the stream of information symbols is organized in blocks, which are precoded through multiplication with the corresponding precoding matrix  $\mathbf{E}_1$  or  $\mathbf{E}_2$ , are appended with the CP, and are then serially transmitted over the asynchronous two-way relay channel after power adjustment. At the receiver front-end of each transceiver, the received signals are turned into blocks of signals, are passed through the self-interference cancelation block, denoted as “SIC”, and are stripped off of the CP. Last but not least, the so-obtained blocks of received signals are linearly transformed (via multiplication by the corresponding post-channel equalizing matrix  $\mathbf{F}_1$  or  $\mathbf{F}_2$ ) with the aim to obtain linear estimates of the transmitted symbol blocks.

At the input of the post-channel equalizer of Transceiver  $q$ , the  $N_s \times 1$  vector of the received signal, denoted as  $\hat{\mathbf{r}}_q$ , is expressed as

$$\hat{\mathbf{r}}_q = \sqrt{P_q} \tilde{\mathbf{H}}(\mathbf{w}) \mathbf{E}_q \tilde{\mathbf{s}}_q + \text{noise}. \quad (9.118)$$

---

<sup>8</sup>These two amplification factors have to be added to the design parameters and have to be obtained optimally through minimizing the total MSE over all design parameters. Having these amplification factors in the design process allows the received symbols be received with gains other than 1.

**FIG. 9.6**

Block diagram of a single-carrier joint pre-channel and post-channel equalization-based two-way relay network [59].

The noise term in Eq. (9.118) is the sum of two noise vectors: the vector of received noises at Transceiver  $q$  and a vector which depends on the relay noises. As shown in Fig. 9.5, this is how the latter noise vector is formed: at each relay, the received noise is multiplied by the relay beamforming weight. This weighted noise then goes through the channel between that relay and Transceiver  $\bar{q}$ , and thus, is delayed and attenuated accordingly. The delayed and attenuated versions of these weighted noises of different relays add up at Transceiver  $\bar{q}$ . Then, this relay noise mixture received at Transceiver  $\bar{q}$  goes through the “S/P” block (i.e., it is turned into a vector), and then, passes through the CP removal block.

Denoting the precoding matrix and channel equalization matrices at Transceiver  $q$  as  $\mathbf{E}_q$  and  $\mathbf{F}_q$ , respectively, the corresponding total MSE minimization problem can be expressed as

$$\begin{aligned} & \min_{\substack{P_1 \geq 0 \\ P_2 \geq 0}} \min_{\mathbf{w}} \min_{\mathbf{E}_1, \mathbf{E}_2} \min_{\mathbf{F}_1, \mathbf{F}_2} \sum_{q=1}^2 \text{MSE}_q(P_{\bar{q}}, \mathbf{F}_q, \mathbf{w}, \mathbf{E}_{\bar{q}}), \\ & \text{subject to } P_T \leq P_T^{\max} \text{ and } \|\mathbf{E}_q\|_F^2 = N_s, \text{ for } q = 1, 2, \end{aligned} \quad (9.119)$$

where the constraint  $\|\mathbf{E}_q\|_F^2 = N_s$  ensures that the transmit power of Transceiver  $q$  is  $P_q$ , for  $q = 1, 2$  while  $\text{MSE}_q(P_{\bar{q}}, \mathbf{F}_q, \mathbf{w}, \mathbf{E}_{\bar{q}})$  is the MSE<sup>9</sup> between the symbols transmitted by Transceiver  $\bar{q}$  and their linear estimates obtained at Transceiver  $q$ , for  $q = 1, 2$ , and can be written as

$$\begin{aligned} \text{MSE}_q(P_{\bar{q}}, \mathbf{F}_q, \mathbf{w}, \mathbf{E}_{\bar{q}}) & \triangleq E\{\|\tilde{\mathbf{s}}_{\bar{q}} - \mathbf{F}_q \mathbf{r}_q\|^2\} \\ & = tr[\mathbf{F}_q \mathbf{R}_q(\mathbf{w}) \mathbf{F}_q^H] - \sqrt{P_{\bar{q}}} tr[\mathbf{F}_q \tilde{\mathbf{H}}(\mathbf{w}) \mathbf{E}_{\bar{q}} + \mathbf{E}_{\bar{q}}^H \tilde{\mathbf{H}}^H(\mathbf{w}) \mathbf{F}_q^H] + N_s. \end{aligned} \quad (9.120)$$

Here,  $\mathbf{R}_q(\mathbf{w})$  represents the correlation matrix of the received signal block at Transceiver  $q$  and can be written, based on Eq. (9.118), as

$$\mathbf{R}_q(\mathbf{w}) = P_{\bar{q}} \tilde{\mathbf{H}}(\mathbf{w}) \mathbf{E}_{\bar{q}} \mathbf{E}_{\bar{q}}^H \tilde{\mathbf{H}}^H(\mathbf{w}) + \sigma^2 (\mathbf{w}^H \mathbf{D}_q \mathbf{w} + 1) \mathbf{I}_{N_s}. \quad (9.121)$$

The optimal value of  $\mathbf{F}_q$  can be obtained by differentiating Eq. (9.120) with respect to  $\mathbf{F}_q$  and equating the derivative to zero. Doing so yields the optimal value of  $\mathbf{F}_q$  as

$$\mathbf{F}_q^o(P_{\bar{q}}, \mathbf{w}, \mathbf{E}_{\bar{q}}) = \sqrt{P_{\bar{q}}} \mathbf{E}_{\bar{q}}^H \tilde{\mathbf{H}}^H(\mathbf{w}) \mathbf{R}_q^{-1}(\mathbf{w}). \quad (9.122)$$

In light of Eqs. (9.96) and (9.122), the optimization problem (9.119) can be simplified. Based on this simplification, the following result is proven in [59].

### RESULT 9.33

The solution to the optimization problem (9.119) is such that only one of the entries of vector of the taps of end-to-end CIR  $\mathbf{h}(\mathbf{w}) = \mathbf{A}\mathbf{w}$  is nonzero.

---

<sup>9</sup>Here, with a slight abuse of notation, one more argument, namely  $\mathbf{E}_{\bar{q}}$  has been added to the list of the arguments of the  $\text{MSE}_q(\cdot, \cdot, \cdot)$ .

Result 9.33 can then be used to prove the following result.

### RESULT 9.34

At the optimum of the optimization problem (9.119), the precoding matrices are unitary, that is, the MSE-optimal values of the precoding matrices satisfy  $\mathbf{E}_q^H \mathbf{E}_q = \mathbf{I}_{N_s}$ , for  $q = 1, 2$ .

According to Result 9.34, one can choose  $\mathbf{E}_q = \mathbf{I}_{N_s}$ , resulting in a single-carrier communication scheme, or  $\mathbf{E}_q = \mathbf{F}^H$  implying a multi-carrier scheme. Both choices result in the same minimum value for the total MSE. The choice  $\mathbf{E}_q = \mathbf{I}_{N_s}$  turns the post-channel equalizer in Eq. (9.122) into the post-channel equalizer in Eq. (9.97). As a result, the remaining design parameters can be obtained using Result 9.26. This conclusion fits squarely with the earlier results regarding the single- and multi-carrier equalization schemes.

Another interesting question to answer is: under a total power constraint what are the sum-rate optimal values of the precoding matrices, the optimal values of the post-channel equalizing matrices, the optimal value of the network beamforming weight vector, and the optimal values of the transceiver transmit powers? Answering this question amounts to solving the following optimization problem:

$$\begin{aligned} & \min_{\substack{P_1 \geq 0 \\ P_2 \geq 0}} \max_{\mathbf{w}} \max_{\mathbf{E}_1, \mathbf{E}_2} \max_{\mathbf{F}_1, \mathbf{F}_2} \sum_{q=1}^2 R_q(P_{\bar{q}}, \mathbf{F}_q, \mathbf{w}, \mathbf{E}_{\bar{q}}), \quad \text{subject to } P_T \leq P_T^{\max} \text{ and } \|\mathbf{E}_q\|_F^2 = N_s, \\ & \text{for } q = 1, 2, \end{aligned} \tag{9.123}$$

where, with small abuse of notation,  $R_q(P_{\bar{q}}, \mathbf{F}_q, \mathbf{w}, \mathbf{E}_{\bar{q}})$  is the achievable rate at Transceiver  $q$ . Recall that Result 9.29 presents the sum-rate-optimal solution to network beamforming problem. This solution results in the largest achievable sum-rate with any precoding or post-channel equalization. As a result, Result 9.29 presents the sum-rate optimal values of the design parameters  $P_1, P_2, \{\mathbf{F}_q\}_{q=1}^2$ , and  $\mathbf{w}$ , as the solution presented in Result 9.26, while the optimal precoders are chosen as  $\mathbf{E}_1 = \mathbf{E}_2 = \mathbf{I}$ .

#### 9.4.3 NETWORKS WITH FREQUENCY-SELECTIVE TRANSCEIVER-RELAY LINKS

In networks with frequency-selective transceiver-relay links, the techniques presented in the previous subsection may not result in satisfactory performance. The reason is that in such networks, using one beamforming weight per relay, each relay may contribute to more than one tap of the end-to-end CIR, due to the frequency-selectivity of transceiver-relay links. Also, due to the time-dispersion nature of the relay-transceiver links, the noise component arrived at the transceivers from each array will be temporally correlated, and thus, the results presented in the previous subsection do not hold.

To combat ISI caused by frequency-selectivity of the transceiver-relay links in two-way relay networks, two competing approaches have been presented in the literature. One approach relies on using OFDM *at all nodes* to “diagonalize” the end-to-end channel, while the second approach relies on FF relaying protocol thereby equalizing the end-to-end channel in a distributed manner.

#### 9.4.3.1 OFDM-based channel equalization

Considering a bidirectional AF-based multi-carrier multi-relay network with frequency-selective transceiver-relay links, the study in [60] considers two different joint power allocation and distributed (relay) beamforming problems.

In the first problem, the aim is to minimize the total power consumed in the entire network, subject to two constraints on the data rates of the two transceivers. In the second problem, the goal is to maximize the sum-rate of the two transceivers subject to a constraint on the total power consumed in the entire network. In both problems, the design parameters are the subcarrier powers at both transceivers and the relay complex beamforming coefficients. It is shown in [60] that the first problem can be tackled using a two-step iterative technique. In the first step, given the subcarrier data rates, the relay beamforming weight vectors and the transceiver power allocation over different subcarriers are obtained using a semi-closed-form. Indeed, this solution relies on the application of the technique presented in Section 9.4.1.1 under Result 9.11, at each subcarrier. In the second step, for fixed relay beamforming weights, the power allocation and the data rates of the transceivers over all subcarriers are calculated by solving a convex optimization problem with a water-filling type of solution. The overall solution iterates between these two optimization problems until convergence is reached.

In the second problem considered in [60], the sum-rate is shown to be *bi-convex* in two vectors, namely the vector of the powers consumed in the entire network over different subcarriers and the vector of the powers of one of the transceivers over different subcarriers. Based on this bi-convexity property of the sum-rate, a two-step iterative algorithm can be used to provide a solution to the sum-rate maximization problem. In the first step, for fixed vector of the total powers consumed in the entire network over different subcarriers, the subproblem of jointly optimal relay beamforming and transceiver power allocation over all subcarriers is solved using a computationally efficient algorithm which relies on the convexity of the subproblem. In the second step, for fixed relay beamforming weight vectors and transceiver power allocations over all subcarriers, the optimal values of the total power consumed in the entire network over all subcarriers are obtained by solving another convex optimization problem. The overall solution iterates between these two optimization problems until convergence is reached. It is worth mentioning that the global optimality cannot be claimed for these iterative techniques.

Assuming a multi-carrier two-way relay network with one relay node, the study in [61] investigates the problem of joint subchannel pairing and power allocation. The approach of [61] relies on maximizing the achievable sum-rate in the network under per node individual power constraints. The challenge in solving this

maximization resides in the fact that this problem is a mixed integer programming problem. Aiming to solve this problem, the study in [61] proposes an iterative algorithm by decomposing the problem into subchannel pairing optimization and joint power allocation optimization, and solving them iteratively. Each subproblem is amenable to computationally efficient solution. Note that for this iterative algorithm, convergence to the global optimal cannot be claimed.

#### **9.4.3.2 Filter-and-forward relaying**

In two-time-slot FF-based two-way relay networks, both transceivers transmit their data simultaneously to the relays in the first time-slot. At each relay, the received signal is processed using an FIR filter, thereby compensating the frequency selectivity of the transceiver-relay channels. The output of the filter is then forwarded to the two transceivers in the second time-slot. The study in [2] presents four different approaches to jointly design the relay FIR filters for such two-way relay networks. The first two approaches assume that the transmit powers of the transceivers are given and fixed. The first of these two approaches aims to minimize the total transmit power of the relays subject to two constraints on the SINR at both transceivers. The second approach designs the relay FIR filters by maximizing the smaller of the two transceivers' SINRs while maintaining the relay transmitted power under certain levels. It is shown in [2] that these two problems can be cast as second-order convex cone programming problems which are amenable to computationally efficient solutions.

The other two design approaches presented in [2] aim to simultaneously obtain the transceivers' transmit powers and the relay FIR filters. The first of these two approaches relies on minimizing the total transmit power consumed in the entire network subject to SINR requirements at the two transceivers, and the second approach maximizes the smaller of the transceivers' SINRs subject to a constraint on the total transmit power consumed in the entire network. Simulation results demonstrate that using an FF relaying strategy can significantly improve the underlying performance measure as compared to the traditional AF relaying approach.

The investigation in [3] considers FF technique, with both FIR and IIR filters at the relays, in conjunction with slicing, linear equalization, or decision-feedback equalization at the transceivers. By assuming slicing at the transceivers and FIR filters at the relays, the design approaches used in [3] are the maximization of the smaller transceivers' SINR subject to a relay transmit power constraint and the minimization of the total relay transmit power subject to QoS constraints. Both problems are shown to be amenable to second-order conic programming formulation, and thus, are efficiently solved.

Assuming linear equalization or decision-feedback equalization at the transceivers, the authors of [3] optimize IIR or FIR filters at the relays by maximizing the smaller SINR and also minimizing the sum-MSE at the equalizer outputs of both transceivers. Leveraging results from FF technique for one-way relaying, the authors establish an upper bound and an achievable lower bound for the max-min problem and an exact solution for the sum-MSE minimization problem. They further show

that the gap between the upper bound and the lower bound for the max-min problem is small, thus their solution performs close to optimal.

#### 9.4.4 MISCELLANEOUS RESULTS

In addition to the previously reviewed results, the list below summarizes two other major research directions related to network beamforming for two-way relay networks.

- The problem of distributed beamforming design for two-way networks with MIMO relays has been considered in numerous published results, see [62–67] and references therein.
- The problem of resource sharing and leasing between two bidirectional relay networks, which employ network beamforming, is studied in [68, 69].

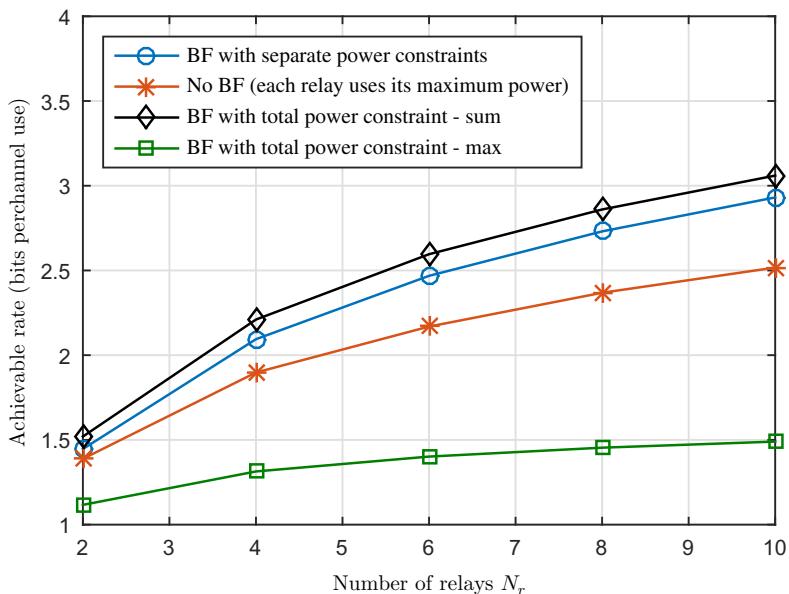
### 9.5 NUMERICAL EXAMPLES

In this section, simulation results are presented with the aim to demonstrate the performance and advantages of network beamforming in one-way and two-way relay networks.

#### 9.5.1 ONE-WAY NETWORK BEAMFORMING

In the context of one-way network beamforming, with perfect synchronization and independent and identically distributed (i.i.d.) flat-fading channels following complex Gaussian distribution with zero-mean and unit-variance., the following four schemes are considered: (1) network beamforming under separate relay power constraints, (2) no beamforming where each relay always uses its maximum power for transmissions, (3) network beamforming under a total relay power constraint where the relay total power constraint is set to be the summation of all relay powers in the separate power case (denoted as “BF with a total relay power constraint—sum”), and (4) network beamforming under a total relay power constraint where the relay total power constraint is set to be the maximum of all relay powers in the separate power case (denoted as “BF with a total relay power constraint—max”). By following the notation in previous sections, let  $Q_r^{\max}$  be the power constraint for Relay  $r$  and  $Q_T^{\max}$  be the total power constraint of all relays. In Scheme 3,  $Q_T^{\max} = \sum_{r=1}^{N_r} Q_r^{\max}$ ; and in Scheme 4,  $Q_T^{\max} = \max_{r \in \{1, \dots, N_r\}} \{Q_r^{\max}\}$ . It is apparent that the performance of Scheme 3 is no worse than that of Scheme 2 due to its larger feasible region for the relay powers, and the performance of Scheme 4 is no better than that of Scheme 2 for the same reason. Schemes 2 and 3 always use the same amount of transmit power regardless of the channel realization. In general, Scheme 1 uses less amount of transmit power than Schemes 2 and 3, and consumes more amount of transmit power than Scheme 4.

Fig. 9.7 shows the achievable rates of networks with single-source single-destination and different numbers of relays. The source transmit power is set to be 10 dB above the noise power. For Scheme 2,  $Q_r^{\max}$  is also set to be 10 dB above

**FIG. 9.7**

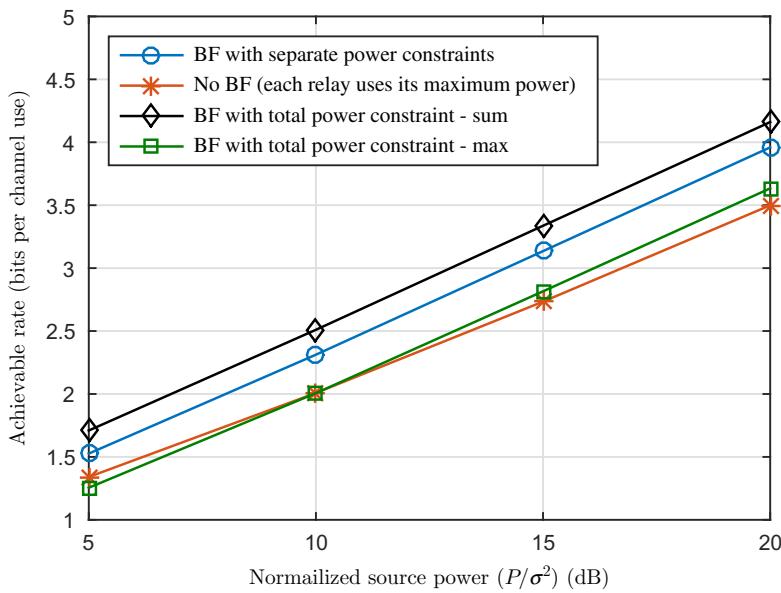
Achievable rate of single source-destination network with different number of relays,  $P/\sigma^2 = 10$  dB.

the noise power. Thus, for Scheme 3,  $Q_T^{\max}/\sigma^2 = 10N_r$ ; and for Scheme 4,  $Q_T^{\max}/\sigma^2 = 10$  dB. It can be seen from the figure that for both separate and total relay power constraint cases, proper beamforming at the relays can largely improve the achievable rate of the relay network and its performance advantage increases as the number of relays increases. The advantage of Scheme 2 over Scheme 4 is due to the  $R$  times higher relay transmit power.

Fig. 9.8 shows the achievable rates of a network with single-source single-destination and different source transmit powers. The number of relays is set as 5. For Scheme 2, the relay power constraints for the five relays are set as  $2P$ ,  $3P/2$ ,  $P/2$ ,  $P/3$ , and  $4P$ , respectively. Thus, for Scheme 3,  $Q_T^{\max} = 19P/3$  and for Scheme 4,  $Q_T^{\max} = 4P$ . It can be seen from the figure that optimal network beamforming designs largely increase the achievable rates of the network for both separate and total relay power constraint cases. With network beamforming, Scheme 4 can achieve higher rates compared to Scheme 3 with considerably lower total transmit power when  $P/\sigma^2$  is larger than 10 dB.

### 9.5.2 TWO-WAY NETWORK BEAMFORMING

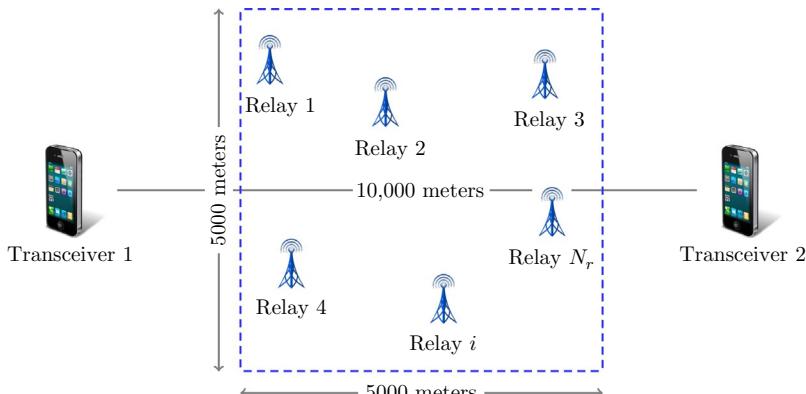
In this subsection, we investigate the numerical performance of the network beamforming algorithm for the single-carrier communication scheme for an asynchronous two-way relay network for different values of  $N_r = 10, 20, 40, 60, 80, 100$  relay

**FIG. 9.8**

Achievable rate of single source-destination network with different source transmit power for  $N_r = 5$  relays.

nodes. To include the path loss and shadowing effects, the channel is modeled based on IEEE 802.16j standard specifications. A path loss factor of 3.8 and a shadowing standard deviation of 8 dB are used to model the large scale fading, and as for the small scale fading, the Rayleigh model is assumed.

As shown in Fig. 9.9, we consider a scenario where two transceivers are 10,000 m apart, and the relay nodes are distributed randomly over an area of 5000 m  $\times$  5000 m,

**FIG. 9.9**

The geometry of the two-way relay network considered.

which is symmetric with respect to the line connecting the two transceivers. Monte Carlo simulation is run for 100 runs, and for each run the total transmit power is varied from 10 dBm to 50 dBm, and the noise power is assumed to be  $-130$  dBm. Furthermore, a symbol rate of 20 Ms/s, or equivalently, a symbol interval of  $T_s = 50$  ns, is assumed. Based on this geometry, the end-to-end CIR will have a delay spread of maximum of  $N = 763$  taps.

Assuming QPSK modulation, Fig. 9.10 compares the BER of the end-to-end two-way communication link versus the total transmit power,  $P_T^{\max}$ , for the network setups with 10, 20, 40, 60, 80, and 100 relay nodes. Also, the relay noise and the receiver noise are modeled as zero-mean unit-variance white Gaussian noise processes. As we expect, Fig. 9.10 shows that increasing the number of relay nodes clearly improves the BER of the networks.

Fig. 9.11 depicts the achievable sum-rate value for different numbers of relay nodes versus the total transmit power over the network. From Fig. 9.11, we notice that in the low-power scenario, namely when the total transmit power is less than 30 dBm, increasing the number of relay nodes is not effective in terms of increasing the sum-rate value. Employing a large number of relay nodes, however, becomes quite advantageous when higher values of total transmit power are allowed.

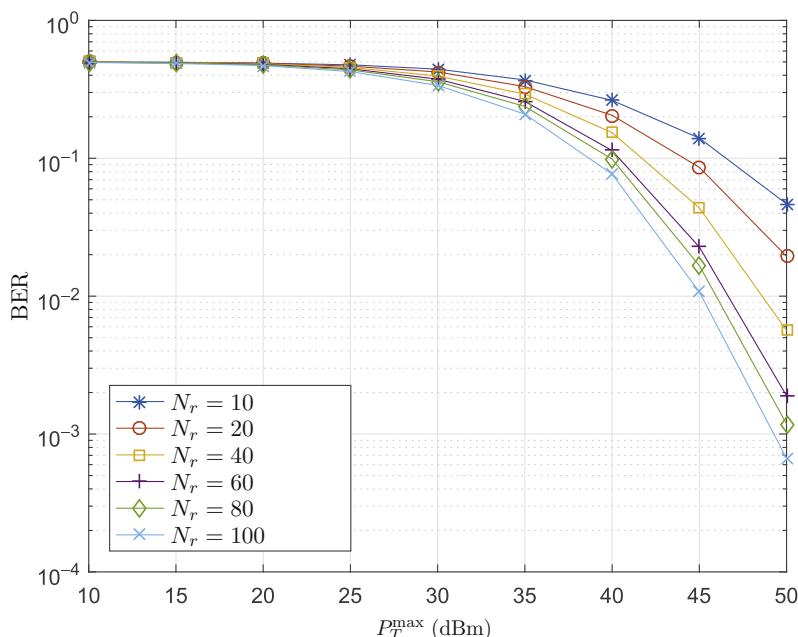
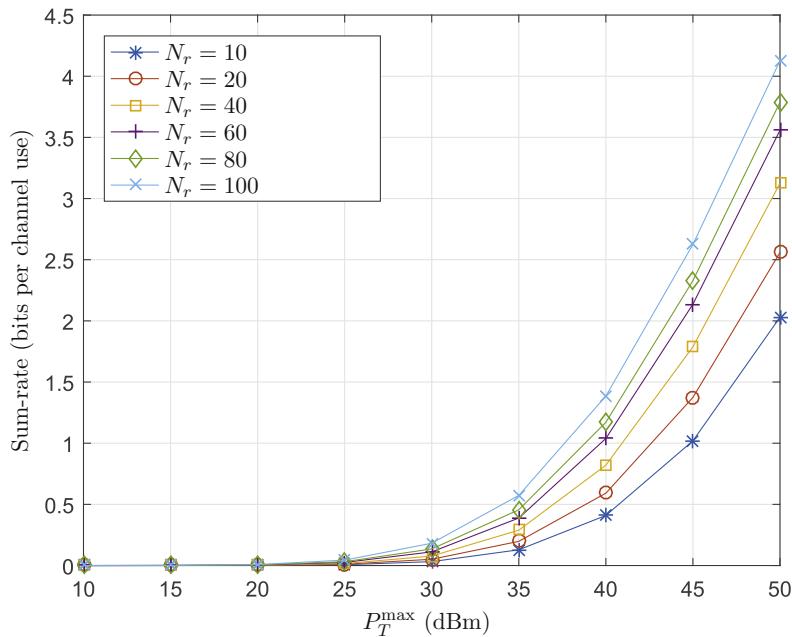


FIG. 9.10

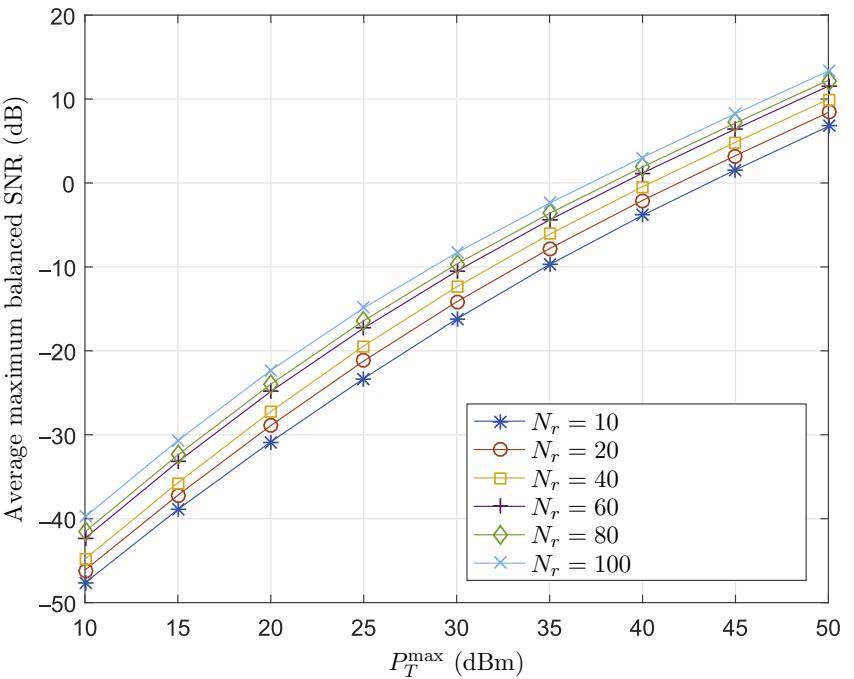
Bit error rate versus the total available transmit power,  $P_T^{\max}$ , for different numbers of relay nodes in a single-carrier/multi-carrier two-way relay network.

**FIG. 9.11**

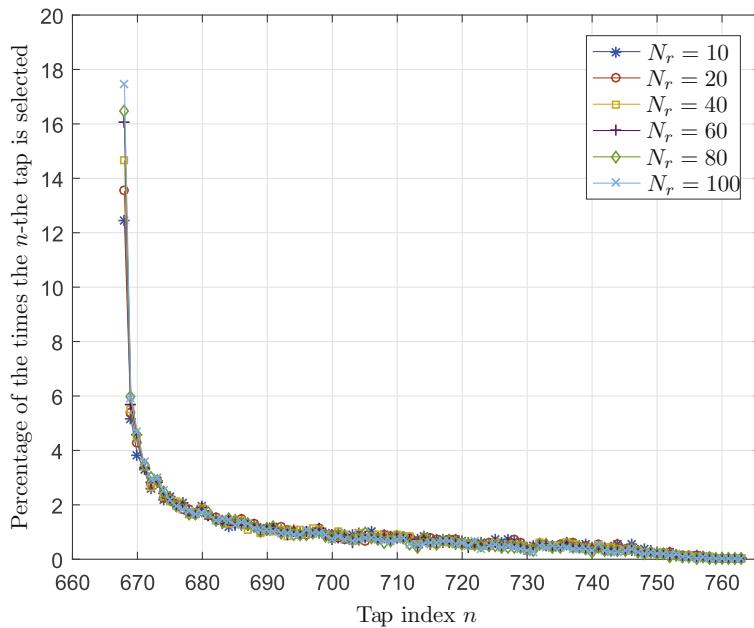
The sum-rate curves versus the total transmit power,  $P_T^{\max}$ , for different numbers of relay nodes in a single-carrier/multi-carrier two-way relay network.

A comparison between the average maximum balanced SNR for different values of relay nodes versus the total transmit power is shown in Fig. 9.12. Average maximum balanced SNR is used as a design criterion to compare the performance of the network when the number of relay nodes changes from 10 relays to 100 relays. As shown in this figure, the average maximum balanced SNR of the network increases gradually as the total power of the network is increased. More importantly, this figure also shows that increasing the number of the relay nodes in the network beyond a certain value does not drastically affect the performance of the network in terms of average maximum balanced SNR.

Fig. 9.13 illustrates the percentage of the times the  $n$ th tap of the end-to-end channel and its associated relays are active, while the other taps remain zero. Fixing the total transmit power at  $P = 60$  dBm, this plot highlights some interesting observations regarding the relay-selection in the studied two-way relay network. Note that tap  $n = 668$  corresponds to those relays which are close to the direct line connecting the two transceivers. That is, these relays contribute to the minimum possible delay between the two transceivers. Based on the considered geometry, the maximum possible delay is  $n = 763$  taps. As can be seen from this figure, the percentage of the times the shortest path is chosen ranges from 12.5% to 17.5%, when the number of the relays changes from 10 to 100. This means that in this figure, in 82.5%–87.5% of the simulation runs selecting the shortest relaying path is not optimal.

**FIG. 9.12**

Average maximum balanced SNR versus total transmit power,  $P_T^{\max}$ , for different numbers of relay nodes in a single-carrier/multi-carrier two-way relay network.

**FIG. 9.13**

Percentage of the selection of the  $n$ th tap of the end-to-end CSI response versus tap index  $n$  plotted for different number of relay nodes, the total transmit power  $P_T = 40$  dBm and 100 runs of simulation.

## 9.6 SUMMARY

This chapter summarizes recent advances in the area of network beamforming in collaborative communication schemes. Over the past decade, a significant volume of research endeavors have been dedicated to this area. Bringing all these results in one chapter does not appear to be feasible. Nevertheless, the authors have aimed to summarize some of the key results appeared in the literature, while striking a balance between comprehensibility and comprehensiveness of the chapter as well as the cohesiveness of the topics. Naturally, the chapter focuses more on the results to which the authors have contributed to and on the results which the authors are aware of at the time of writing this chapter.

The chapter starts off with [Section 9.1](#) presenting a model for an end-to-end relay channel. This model is used throughout the rest of chapter. [Section 9.2](#) provides an overview on network beamforming for one-way relaying schemes under different assumptions, for different scenarios, and using different design criteria. Both single- and multi-user cases are presented.

[Section 9.3](#) focuses on three types of two-way (bi-directional) relay networks: (1) synchronous two-way relay networks, (2) asynchronous two-relay networks,

and (3) networks with frequency-selective transceiver-relay links. Different design criteria as well as the corresponding solutions are presented. The relationships between different design approaches are discussed.

The authors believe that what has been presented in this chapter is the tip of the iceberg, compared to the volume of the published results in the area of network beamforming and decentralized signal processing for cooperative networks. Our hope is that readers find this chapter helpful in shaping new research directions.

---

## REFERENCES

- [1] H. Chen, A.B. Gershman, S. Shahbazpanahi, Filter-and-forward distributed beamforming in relay networks with frequency selective fading, *IEEE Trans. Signal Process.* 58 (3) (2010) 1251–1262.
- [2] H. Chen, S. ShahbazPanahi, A.B. Gershman, Filter-and-forward distributed beamforming for two-way relay networks in frequency selective channels, *IEEE Trans. Signal Process.* 60 (2012) 1927–1941.
- [3] Y.-W. Liang, A. Ikhlef, W. Gerstacker, R. Schober, Cooperative filter-and-forward beamforming for frequency-selective channels with equalization, *IEEE Trans. Wireless Commun.* 10 (2011) 228–239.
- [4] Y.-W. Liang, R. Schober, Cooperative amplify-and-forward beamforming with multiple multi-antenna relays, *IEEE Trans. Commun.* 59 (2011) 2605–2615.
- [5] Y.-W. Liang, A. Ikhlef, W. Gerstacker, R. Schober, Two-way filter-and-forward beamforming for frequency-selective channels, *IEEE Trans. Wireless Commun.* 10 (2011) 4172–4183.
- [6] J. Mirzaee, S. ShahbazPanahi, R. Vahidnia, Sum-rate maximization for active channels, *IEEE Signal Process Lett.* 20 (2013) 771–774.
- [7] J. Mirzaei, S. ShahbazPanahi, Sum-rate maximization for active channels with unequal subchannel noise powers, *IEEE Trans. Signal Process.* 62 (2014) 4187–4198.
- [8] P. Abbasi-Saei, S. Shahbazpanahi, Sum-rate maximization for two-way active channels, *IEEE Trans. Signal Process.* 64 (2016) 1369–1382.
- [9] A. Kiani, S. Shahbazpanahi, Sum-rate maximization for two-way active channels with unequal subcarrier noise powers, *IEEE Trans. Signal Process.* (2017) (submitted for publication).
- [10] P. Larsson, Large-scale cooperative relaying network with optimal coherent combining under aggregate relay power constraints, in: Proc. Future Tele. Conf. 2003.
- [11] Y. Jing, H. Jafarkhani, Network beamforming using relays with perfect channel information, *IEEE Trans. Inf. Theory* 55 (2009) 2499–2517.
- [12] G. Zheng, K.-K. Wong, A. Paulraj, B. Ottersten, Collaborative-relay beamforming with perfect CSI: optimum and distributed implementation, *IEEE Signal Process Lett.* 16 (2009) 257–260.
- [13] Y. Hao, Y. Jing, S. ShahbazPanahi, Energy efficient network beamforming design using power-normalized SNR, *IEEE Trans. Wireless Commun.* 13 (5) (2014) 2756–2769.
- [14] Y. Jing, H. Jafarkhani, Single and multiple relay selection schemes and their achievable diversity orders, *IEEE Trans. Wireless Commun.* 8 (2009) 1414–1423.
- [15] E. Koyuncu, Y. Jing, H. Jafarkhani, Distributed beamforming in wireless relay networks with quantized feedback, *IEEE J. Sel. Areas Commun.* 26 (2008) 1429–1439.

- [16] V. Havary-Nassab, S. Shahbazpanahi, A. Grami, Z.Q. Luo, Distributed beamforming for relay networks based on second-order statistics of the channel state information, *IEEE Trans. Signal Process.* 56 (2008) 4306–4316.
- [17] J. Li, A.P. Petropulu, H.V. Poor, Cooperative transmission for relay networks based on second-order statistics of channel state information, *IEEE Trans. Signal Process.* 59 (2011) 1280–1291.
- [18] N. Khajehnouri, A.H. Sayed, Distributed MMSE relay strategies for wireless sensor networks, *IEEE Trans. Signal Process.* 55 (2007) 3336–3348.
- [19] J. Choi, MMSE-based distributed beamforming in cooperative relay networks, *IEEE Trans. Commun.* 59 (2011) 1346–1356.
- [20] J. Choi, Distributed beamforming using a consensus algorithm for cooperative relay networks, *IEEE Commun. Lett.* 15 (2011) 368–370.
- [21] L. Zhang, W. Liu, A.U. Quddus, M. Dianati, R. Tafazolli, Adaptive distributed beamforming for relay networks based on local channel state information, *IEEE Trans. Signal Inf. Process. Networks* 1 (2015) 117–128.
- [22] D.H.N. Nguyen, H.H. Nguyen, Power allocation in wireless multi-user multi-relay networks with distributed beamforming, *IET Commun.* 5 (14) (2011) 2040–2051.
- [23] Q. Wang, Y. Jing, Power allocation and sum-rate analysis for multi-user multi-relay networks, in: 2013 IEEE 78th Vehicular Technology Conference (VTC Fall), 2013, pp. 1–5.
- [24] A. Ramezani-Kebrya, M. Dong, B. Liang, G. Boudreau, R. Casselman, Per-relay power minimization for multi-user multi-channel cooperative relay beamforming, *IEEE Trans. Wireless Commun.* 15 (2016) 3187–3198.
- [25] U. Rashid, H.D. Tuan, H.H. Nguyen, Maximin relay beamforming in multi-user amplify-forward wireless relay networks, in: 2012 IEEE Wireless Communications and Networking Conference (WCNC), 2012, pp. 3008–3012.
- [26] U. Rashid, H.D. Tuan, H.H. Nguyen, Relay beamforming designs in multi-user wireless relay networks based on throughput maximin optimization, *IEEE Trans. Commun.* 61 (2013) 1739–1749.
- [27] E. Che, H.D. Tuan, H.H. Nguyen, Joint optimization of cooperative beamforming and relay assignment in multi-user wireless relay networks, *IEEE Trans. Wireless Commun.* 13 (2014) 5481–5495.
- [28] C.C. Chen, C.S. Tseng, J. Denis, C. Lin, Adaptive distributed beamforming for amplify-and-forward relay networks: convergence analysis, *IEEE Trans. Wireless Commun.* 13 (2014) 4167–4178.
- [29] E. Koyuncu, H. Jafarkhani, Distributed beamforming in wireless multi-user relay-interference networks with quantized feedback, *IEEE Trans. Inf. Theory* 58 (2012) 4538–4576.
- [30] S. Fazeli-Dehkordy, S. Shahbazpanahi, S. Gazor, Multiple peer-to-peer communications using a network of relays, *IEEE Trans. Signal Process.* 57 (8) (2009) 3053–3062.
- [31] N. Chatzipanagiotis, Y. Liu, A. Petropulu, M.M. Zavlanos, Distributed cooperative beamforming in multi-source multi-destination clustered systems, *IEEE Trans. Signal Process.* 62 (2014) 6105–6117.
- [32] B. Mahboobi, E.S.-Nasab, M. Ardebilipour, Outage probability based robust distributed beam-forming in multi-user cooperative networks with imperfect CSI, *Wirel. Pers. Commun.* 77 (3) (2014) 1629–1658.
- [33] B. Mahboobi, M. Ardebilipour, A. Kalantari, E. Soleimani-Nasab, Robust cooperative relay beamforming 2 (2013) 399–402.

- [34] M.A.M. Sadr, M. Ahmadian-Attari, B. Mahboobi, Low-complexity robust relay optimisation for multiple peer-to-peer beamforming: a safe tractable approximation approach, *IET Commun.* 9 (16) (2015) 1968–1979.
- [35] M.A.M. Sadr, B. Mahboobi, S. Mehrizi, M.A. Attari, M. Ardebilipour, Stochastic robust collaborative beamforming: non-regenerative relay, *IEEE Trans. Commun.* 64 (2016) 947–958.
- [36] D. Ponukumati, F. Gao, C. Xing, Robust peer-to-peer relay beamforming: a probabilistic approach, *IEEE Commun. Lett.* 17 (2013) 305–308.
- [37] T. Wang, B.P. Ng, M.H. Er, Frequency-domain approach to relay beamforming with adaptive decision delay for frequency-selective channels, *IEEE Trans. Signal Process.* 61 (2013) 5563–5577.
- [38] A. Schad, H. Chen, A. Gershman, S. Shahbazpanahi, Filter-and-forward multiple peer-to-peer beamforming in relay networks with frequency selective channels, in: *Proc. ICASSP'10*, 2010, pp. 3246–3249.
- [39] N. Bornhorst, M. Pesavento, Filter-and-forward beamforming with adaptive decoding delays in asynchronous multi-user relay networks, *Signal Process.* 109 (3) (2015) 132–147.
- [40] H.H. Kha, H.D. Tuan, H.H. Nguyen, Joint optimization of source power allocation and cooperative beamforming for SC-FDMA multi-user multi-relay networks, *IEEE Trans. Commun.* 61 (2013) 2248–2259.
- [41] M. Zaeri-Amirani, S. Shahbazpanahi, T. Mirfakhraie, K. Ozdemir, Performance trade-offs in amplify-and-forward bidirectional network beamforming, *IEEE Trans. Signal Process.* 60 (8) (2012) 4196–4209.
- [42] V. Havary-Nassab, S. Shahbazpanahi, A. Grami, Optimal distributed beamforming for two-way relay networks, *IEEE Trans. Signal Process.* 58 (3) (2010) 1238–1250.
- [43] S. Shahbazpanahi, M. Dong, A semi-closed-form solution to optimal distributed beamforming for two-way relay networks, *IEEE Trans. Signal Process.* 60 (3) (2012) 1511–1516.
- [44] S. Talwar, S. Shahbazpanahi, A total power minimization approach to relay selection for two-way relay networks, in: *Conference Record of the Forty Sixth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, 2012, 2012, pp. 2001–2005.
- [45] Y. Jing, S. Shahbazpanahi, Max-min optimal joint power control and distributed beamforming for two-way relay networks under per-node power constraints, *IEEE Trans. Signal Process.* 60 (2012) 6576–6589.
- [46] M.O. Hasna, M.-S. Alouini, Optimal power allocation for relayed transmissions over Rayleigh-fading channels, *IEEE Trans. Wireless Commun.* 3 (2004) 1999–2004.
- [47] Y. Zhao, R. Adve, T.J. Lim, Improving amplify-and-forward relay networks: Optimal power allocation versus selection, *IEEE Trans. Wireless Commun.* 6 (2007) 3114–3123.
- [48] R. Vahidnia, S. ShahbazPanahi, Single-carrier equalization for asynchronous two-way relay networks, *IEEE Trans. Signal Process.* 62 (2014) 5793–5808.
- [49] R. Vahidnia, S. Shahbazpanahi, A. Minasian, Pre-channel equalization and distributed beamforming in asynchronous single-carrier bi-directional relay networks, *IEEE Trans. Signal Process.* 64 (2016) 3968–3983.
- [50] S. ShahbazPanahi, M. Dong, Achievable rate region under joint distributed beamforming and power allocation for two-way relay networks, *IEEE Trans. Wireless Commun.* 11 (2012) 4026–4037.
- [51] S. Talwar, Y. Jing, S. Shahbazpanahi, Joint relay selection and power allocation for two-way relay networks, *IEEE Signal Process Lett.* 18 (2) (2011) 91–94.
- [52] M. Dong, S. Shahbazpanahi, Optimal spectrum sharing and power allocation for OFDM-based two-way relaying, in: *Proc. ICASSP'10*, 2010, pp. 3310–3313.

- [53] R. Vahidnia, S. Shahbazpanahi, Multi-carrier asynchronous bi-directional relay networks: joint subcarrier power allocation and network beamforming, *IEEE Trans. Wireless Commun.* 12 (2013) 3796–3812.
- [54] Z. Wang, G.B. Giannakis, Wireless multicarrier communications, *IEEE Signal Process. Mag.* 17 (3) (2000) 29–48.
- [55] J. Mirzaei, S. ShahbazPanahi, On achievable SNR region for multi-user multi-carrier asynchronous bidirectional relay networks, *IEEE Trans. Wireless Commun.* 14 (2015) 3219–3230.
- [56] R. AliHemmati, S. Shahbazpanahi, Sum-rate optimal network beamforming and subcarrier power allocation for multi-carrier asynchronous two-way relay networks, *IEEE Trans. Signal Process.* 63 (2015) 4129–4143.
- [57] M. Askari, S. ShahbazPanahi, Sum-rate optimal network beamforming and power allocation for single-carrier asynchronous bidirectional relay networks, *IEEE Access* 5 (2017) 13699–13711.
- [58] S. Bastanirad, S. Shahbazpanahi, A. Grami, A total power minimization approach to optimal network beamforming and power allocation in single-carrier asynchronous two-way relay networks, *IEEE Trans. Wireless Commun.* (2017) (submitted for review).
- [59] F. Eshaghian Dorcheh, S. ShahbazPanahi, Jointly optimal pre- and post-channel equalization and distributed beamforming in asynchronous bi-directional relay networks, *IEEE Trans. Signal Process.* 65 (2017) 4593–4608.
- [60] R. AliHemmati, S. Shahbazpanahi, M. Dong, Joint spectrum sharing and power allocation for OFDM-based two-way relaying, *IEEE Trans. Wireless Commun.* 14 (2015) 3294–3308.
- [61] M. Chang, M. Dong, F. Zuo, S. ShahbazPanahi, Joint subchannel pairing and power allocation in multichannel MABC-based two-way relaying, *IEEE Trans. Wireless Commun.* 15 (2016) 620–632.
- [62] K.-J. Lee, H. Sung, E. Park, I. Lee, Joint optimization for one and two-way MIMO AF multiple-relay systems, *IEEE Trans. Wireless Commun.* 9 (12) (2010) 3671–3681.
- [63] R. Vaze, R.W. Heath, Capacity scaling for MIMO two-way relaying, in: Proc. IEEE International Symposium on Information Theory, (ISIT'07), Nice, 2007, pp. 1451–1455.
- [64] D. Gunduz, A. Goldsmith, H.V. Poor, MIMO two-way relay channel: Diversity-multiplexing tradeoff analysis, in: Proc. Asilomar Conf. Signals Syst. Comput. (ASILOMAR'08), Pacific Grove, CA, 2008, pp. 1474–1478.
- [65] H.H. Kha, H.D. Tuan, H.H. Nguyen, H.H.M. Tam, Joint design of user power allocation and relay beamforming in two-way MIMO relay networks, in: Proc. Int. Conf. Signal Process. Commun. Syst., Gold Coast, Australia, 2013, pp. 1–6.
- [66] A. Alsharoa, H. Ghazzai, M.S. Alouini, Energy efficient design for MIMO two-way AF multiple relay networks, in: Proc. 2014 IEEE Wireless Communications and Networking Conference (WCNC), Istanbul, 2014, pp. 1007–1011.
- [67] R. Rahimi, S. ShahbazPanahi, A two-way network beamforming approach based on total power minimization with symmetric relay beamforming matrices, *IEEE Access* 5 (2017) 12458–12474.
- [68] A. Gavili, S. ShahbazPanahi, Optimal spectrum leasing and resource sharing in two-way relay networks, *IEEE Trans. Signal Process.* 62 (2014) 5030–5045.
- [69] A. Gavili, S. Shahbazpanahi, Optimal resource sharing and network beamforming in multi-carrier bidirectional relay networks, *IEEE Trans. Signal Process.* 63 (2015) 6354–6367.

# Transmit beamforming for simultaneous wireless information and power transfer 10

Liang Liu\*, Jie Xu<sup>†</sup>, Rui Zhang<sup>‡</sup>

*University of Toronto, Toronto, ON, Canada*<sup>\*</sup> *Guangdong University of Technology, Guangzhou, Guangdong, China*<sup>†</sup> *National University of Singapore, Singapore*<sup>‡</sup>

## 10.1 INTRODUCTION

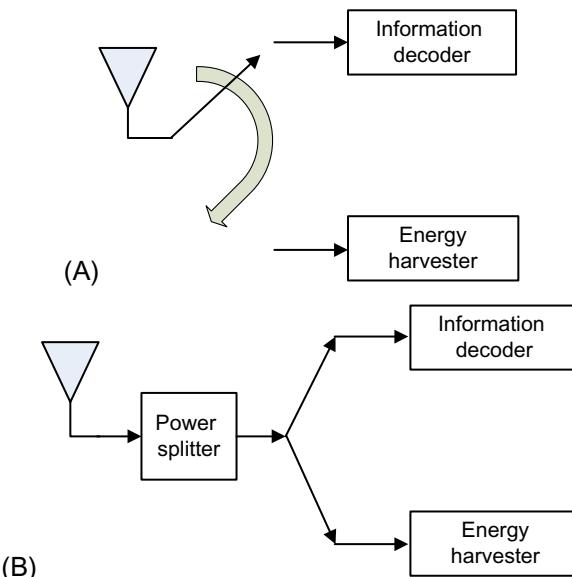
Radio frequency (RF) transmission enabled wireless power transfer (WPT) is a cost-effective solution to power energy-constrained wireless networks (e.g., sensor networks), where dedicated energy transmitters are deployed to broadcast RF signals to charge low-power electric devices (e.g., sensors and RF identification (RFID) tags). Unlike other commonly used energy harvesting sources such as solar and wind that are intermittent and unreliable, RF-based WPT is able to provide continuous and controllable power supply, and thus is applicable to more energy-demanding applications [1]. On the other hand, RF signals have been widely used in wireless communications as the carrier for wireless information transfer (WIT) for several decades. Traditionally, WPT and WIT have been investigated as two separate lines of research. As they are both enabled by RF signals transmission, a practical question thus arises that whether we can utilize RF signals more efficiently for both WPT and WIT at the same time. This question has been recently addressed with the invention and investigation of a new technique called simultaneous wireless information and power transfer (SWIPT), which has attracted rapidly increasing attention in research. SWIPT successfully unified the research in WIT and WPT, and opened an exciting new direction for their joint investigation.

### 10.1.1 PRACTICAL SWIPT RECEIVER

The basic idea of SWIPT was first proposed in [2, 3] by considering a single-antenna point-to-point channel, where the trade-off between the achievable rate for WIT and the received energy for WPT is investigated from an information-theoretic perspective. In these initial studies, the authors assumed that the single-antenna receiver can utilize the same received RF signals for both information decoding (ID) and energy

harvesting (EH) at the same time without any loss. However, this assumption is difficult to realize in practice. This is due to the fact that existing information receivers (IRs) and energy receivers (ERs) are separately designed with distinct circuit structures, and as a result, each of them cannot be used to decode information as well as harvest energy at the same time. To circumvent this practical difficulty, various practical receiver architectures for SWIPT have been proposed in [4, 5]. Among them, two basic receiver structures have been widely adopted in the literature.

- “*Time-Switching (TS)*” receiver: As shown in Fig. 10.1A, the TS receiver switches between an information decoder and an energy harvester over time. The TS scheme is the simplest way to implement SWIPT by using off-the-shelf commercially available circuits for ID and EH, respectively. For TS-based SWIPT receivers, it is crucial to determine their operation modes (ID or EH) over time based on their communication and energy requirements, as well as the channel conditions. For the receiver with multiple antennas, applying TS to each of the antennas independently leads to a low-complexity SWIPT receiver called “antenna switching (AS)” [4].
- “*Power-Splitting (PS)*” receiver: As shown in Fig. 10.1B, the PS receiver splits its received signal into two portions: one for ID and the other for EH. For PS receivers, it is important to determine the power splitting ratio at each antenna to balance the rate-energy trade-off between the ID and EH receivers. Note that TS or AS receiver can be regarded as a special and low-complexity realization of PS



**FIG. 10.1**

An illustration of the TS and PS receivers. (A) TS receiver. (B) PS receiver.

receiver with only binary (0 or 1) power splitting ratio at each receiving antenna; nevertheless, they are implemented by different hardware circuits (time switcher versus power splitter) in practice.

### 10.1.2 MULTIANTENNA SWIPT

Besides the challenge in the receiver design for SWIPT, how to enhance the efficiency of concurrent WIT and WPT to overcome the significant power loss of RF signal propagation over long distances is another important issue for the practical design of SWIPT. Motivated by the great success of multiantenna techniques in wireless communications [6], multiantenna enabled transmit beamforming is also a promising solution for SWIPT. With multiantenna transmit beamforming, the transmitter can send one or more beams to direct the information and/or energy signals more efficiently to the information and energy receivers for WIT and WPT, respectively, and also minimize their co-channel interference that is harmful to the IRs.

In multiantenna WIT systems, transmit beamforming has been widely adopted as an efficient way to improve the communication rate and reliability [7]. Over the last two decades, there have been extensive studies on the optimal transmit beamforming design for wireless communication systems, some examples of which are briefly discussed as follows. In the broadcast channel consisting of one multiantenna transmitter and many single-antenna receivers, the so-called power minimization problem, signal-to-interference-plus-noise ratio (SINR) balancing problem, as well as the weighted sum-rate maximization problem are widely studied in the literature. Specifically, for the power minimization problem, the transmit beamforming vectors are designed to minimize the total transmit power subject to the users' minimum SINR constraints. It is shown that this problem can be reformulated as a second-order cone program (SOCP) and thus efficiently solved by the powerful convex optimization technique [8]. Moreover, this problem can also be solved based on the alternative approach of uplink-downlink duality [9]. For the SINR balancing problem, the objective of the beamforming design is to maximize the minimum SINR among all the users subject to the total transmit power constraint. This problem is efficiently solved by the nonnegative matrix theory [10]. Last but not least, various beamforming design algorithms are proposed to maximize the weighted sum-rate of all the users subject to the total transmit power constraint assuming linear precoding at the multi-antenna transmitter [11, 12].

Similarly, in multiantenna WPT systems, transmit beamforming has been applied to combat the severe signal power loss over long distance to improve the energy transfer efficiency. It has been shown in [13] that transmitting with only one single energy beam to all the ERs is sufficient to maximize their weighted sum energy harvested, which is in sharp contrast to the optimal transmit beamforming in WIT systems, where sending multiple information beams is generally required. Moreover, the joint channel acquisition and energy beamforming design is investigated in [14, 15] to maximize the energy harvested

by all the users based on a practical energy feedback framework exploiting the analytic center cutting plane method in convex optimization.

Although the information beamforming design for WIT and energy beamforming design for WPT are well studied separately, a joint optimization of the information and energy beamforming in TS or PS-based SWIPT systems is a new research problem of both theoretical and practical significance, which has been pursued recently. In this paper, we provide an overview on recent advances in joint information and energy beamforming design for SWIPT systems.

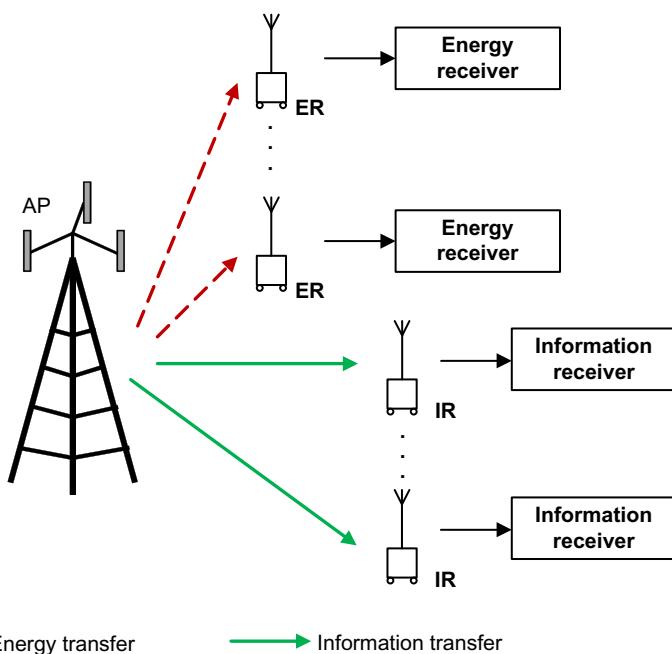
The remainder of this paper is organized as follows. [Section 10.2](#) introduces a multiuser multiple-input single-output (MISO) broadcast channel model for SWIPT with separate or co-located IR and ER receivers, and presents the optimal beamforming design in each case. In particular, for the case of separate IRs and ERs, we also consider the communication security problem in SWIPT where ERs may eavesdrop the signals for IRs, and present the beamforming solution to overcome this problem from a physical-layer perspective. [Section 10.3](#) extends the discussions to other SWIPT system setups. Finally, [Section 10.4](#) concludes this paper.

---

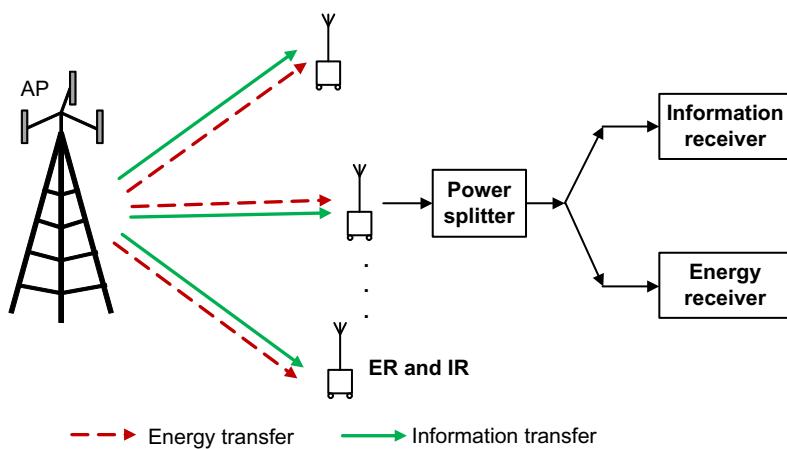
## 10.2 JOINT INFORMATION AND ENERGY BEAMFORMING DESIGN FOR SWIPT

In this section, we consider a MISO point-to-multipoint (broadcast channel) SWIPT system consisting of one access point (AP) equipped with  $M > 1$  antennas and  $K$  users each equipped with one single antenna, denoted by the set  $\mathcal{K} = \{1, \dots, K\}$ . We consider a narrow-band flat-fading channel model, and the channel from the AP to user  $k$  is denoted by  $\mathbf{h}_k \in \mathbb{C}^{M \times 1}$ ,  $\forall k$ . To characterize the fundamental performance limit of transmit beamforming for SWIPT, it is assumed that the AP perfectly knows its channels to all the users, and thus is able to design its transmit signals to meet the ID and EH requirements of the users.

In our considered system, the AP broadcasts RF signals to carry both information and energy to all  $K$  users. Since each single-antenna receiver cannot decode information and harvest energy from the same received RF signal, it is assumed that either the TS or PS receiver in [Fig. 10.1](#) is used at each user. To characterize the SWIPT performance under different receiver structures, we consider two scenarios where all the users use the same type of receivers, which are either TS or PS-based. In particular, [Fig. 10.2](#) illustrates the case with all TS receivers, where at any particular time a subset of the  $K$  users decode information from their received signals, while the other users harvest energy from their received signals. We refer to this setup as SWIPT systems with separate IRs and ERs. On the other hand, [Fig. 10.3](#) shows the case with all PS receivers, where each of the  $K$  users can decode information and harvest energy at the same time by splitting the received signal for ID and EH, respectively. We thus refer to this case as SWIPT systems with co-located IRs and ERs.

**FIG. 10.2**

A MISO SWIPT system with separate IRs and ERs.

**FIG. 10.3**

A MISO SWIPT system with co-located IRs and ERs.

In the following three subsections, we will address three fundamental beamforming design problems under the above two setups. Specifically, we consider first the SWIPT system with separate IRs and ERs in Fig. 10.2. Due to different power sensitivities of the IRs and ERs, a user location based transmission scheduling is proposed in Section 10.2.1, where the users that are close to the AP are scheduled to harvest energy, and the others that are more distant away from the AP are scheduled to receive information. Under this scheduling strategy, the joint information and energy beamforming design is formulated to maximize the weighted sum-energy harvested by all the ERs subject to the minimum SINR constraints for the IRs as well as the total transmit power constraint at the AP. Next, in Section 10.2.2, the physical-layer security for the IRs in SWIPT is studied. Since the ERs are deployed much closer to the AP, they can easily eavesdrop the information for the IRs from their received signals broadcast by the AP. A novel artificial noise (AN) assisted beamforming design is proposed to achieve secrete WIT to the IRs and efficient WPT to the ERs simultaneously. Last, we consider the PS-based SWIPT system with users of co-located IR and ER in Fig. 10.3. In this case, each user has its individual SINR requirement for the IR and harvested energy requirement for the ER, which need to be satisfied at the same time by properly designing the power splitting ratio between the IR and ER. Thus, the joint design of transmit beamforming at the AP and power splitting ratios at the users is considered in Section 10.2.3, with the objective of minimizing the transmit power at the AP subject to each user's minimum SINR and harvested energy constraints.

### 10.2.1 BEAMFORMING DESIGN FOR SWIPT SYSTEM WITH SEPARATE IRS AND ERS

Consider the SWIPT system with separate IRs and ERs in Fig. 10.2. In practice, EH and ID circuits operate with different received power requirements. For example, ERs may require their received power to be above 0.1 mW or  $-10$  dBm, while IRs can operate with a much lower received power even below  $-50$  dBm. Due to the severe wireless signal power loss over distance, the operating range of ERs is often much shorter than that of IRs. To address this practical issue, we consider a receiver location based transmission scheme, where the ERs are deployed sufficiently close to the AP, while the IRs can be located more distant from the AP. Notice that the proposed transmission scheme potentially solves the mismatched power issue for IRs and ERs, and thus makes the SWIPT system practically feasible. Also note that the location-based transmission should be designed in practice by taking into account the potential mobility of receivers to ensure certain fairness in energy and information transmitted to them over time. Under this setup, we aim to jointly design the beamforming weights and power allocation at the transmitter to optimally balance the performance trade-off among different IRs and ERs. More details are given as follows.

### 10.2.1.1 System model

Among the  $K$  users, it is assumed that  $K_I$  users are scheduled to be IRs and the remaining  $K_E = K - K_I$  users are ERs, which are denoted by the sets  $\mathcal{K}_I = \{1, \dots, K_I\}$  and  $\mathcal{K}_E = \{K_I + 1, \dots, K\}$ , respectively. Specifically, we consider linear precoding at the transmitter for SWIPT and each IR/ER is assigned with one dedicated information/energy beam without loss of generality. Hence, the transmitted signal from the AP is given by

$$\mathbf{x} = \sum_{i \in \mathcal{K}_I} \mathbf{v}_i s_i^{\text{ID}} + \sum_{j \in \mathcal{K}_E} \mathbf{w}_j s_j^{\text{EH}}, \quad (10.1)$$

where  $\mathbf{v}_i \in \mathbb{C}^{M \times 1}$  and  $\mathbf{w}_j \in \mathbb{C}^{M \times 1}$  are the beamforming vectors for IR  $i$  and ER  $j$ , while  $s_i^{\text{ID}}$  and  $s_j^{\text{EH}}$  are the information-bearing signal for IR  $i$  and energy-carrying signal for ER  $j$ , respectively. For information signals, Gaussian inputs are assumed, i.e.,  $s_i^{\text{ID}}$ 's are independent and identically distributed (i.i.d.) circularly symmetric complex Gaussian (CSCG) random variables with zero mean and unit variance, denoted by  $s_i^{\text{ID}} \sim \mathcal{CN}(0, 1)$ ,  $\forall i \in \mathcal{K}_I$ . For energy signals, since  $s_j^{\text{EH}}$  carries no information, it can be any arbitrary random signal provided that its power spectral density satisfies certain regulations on microwave radiation. Without loss of generality, we assume that  $s_j^{\text{EH}}$ 's are independent white sequences from an arbitrary distribution with  $\mathbb{E}(|s_j^{\text{EH}}|^2) = 1$ ,  $\forall j \in \mathcal{K}_E$ . Suppose that the AP has a transmit sum-power constraint  $P$ ; from Eq. (10.1) we thus have  $\mathbb{E}(\mathbf{x}^H \mathbf{x}) = \sum_{i \in \mathcal{K}_I} \|\mathbf{v}_i\|^2 + \sum_{j \in \mathcal{K}_E} \|\mathbf{w}_j\|^2 \leq P$ .

The discrete-time baseband signal at IR  $i \in \mathcal{K}_I$  is given by

$$y_i^{\text{ID}} = \mathbf{h}_i \mathbf{x} + z_i, \quad \forall i \in \mathcal{K}_I, \quad (10.2)$$

where  $z_i \sim \mathcal{CN}(0, \sigma_i^2)$  is the i.i.d. Gaussian noise at IR  $i$ . With linear transmit precoding, each IR is potentially interfered with by all other unintended information beams and energy beams. Since energy beams carry no information but instead pseudorandom signals which can be assumed to be known at both the AP and each IR prior to data transmission, their resulting interference can be canceled at each IR if this additional operation is implemented. We thus consider two types of IR, namely Type I and Type II IRs, which do not possess and possess the capability of canceling the interference due to energy signals, respectively. Therefore, for IR  $i$  of Type I or Type II, the corresponding SINR is accordingly expressed as

$$\text{SINR}_i^{(I)} = \frac{|\mathbf{h}_i \mathbf{v}_i|^2}{\sum_{k \neq i, k \in \mathcal{K}_I} |\mathbf{h}_i \mathbf{v}_k|^2 + \sum_{j \in \mathcal{K}_E} |\mathbf{h}_i \mathbf{w}_j|^2 + \sigma_i^2}, \quad \forall i \in \mathcal{K}_I, \quad (10.3)$$

$$\text{SINR}_i^{(II)} = \frac{|\mathbf{h}_i \mathbf{v}_i|^2}{\sum_{k \neq i, k \in \mathcal{K}_I} |\mathbf{h}_i \mathbf{v}_k|^2 + \sigma_i^2}, \quad \forall i \in \mathcal{K}_I. \quad (10.4)$$

On the other hand, for wireless energy transfer, due to the broadcast property of wireless channels, the energy carried by all information and energy beams, i.e., both  $\mathbf{v}_i$ 's

and  $\mathbf{w}_j$ 's, can be harvested at each ER. As a result, the harvested power for ER  $j \in \mathcal{K}_{\mathcal{E}}$ , denoted by  $Q_j$ , is proportional to the total power received [4], i.e.,

$$Q_j = \zeta \left( \sum_{k \in \mathcal{K}_{\mathcal{I}}} |\mathbf{h}_j \mathbf{v}_k|^2 + \sum_{k \in \mathcal{K}_{\mathcal{E}}} |\mathbf{h}_j \mathbf{w}_k|^2 \right), \quad \forall j \in \mathcal{K}_{\mathcal{E}}, \quad (10.5)$$

where  $0 < \zeta \leq 1$  denotes the energy harvesting efficiency.

### 10.2.1.2 Problem formulation

Our aim is to maximize the weighted sum-power transferred to all ERs subject to individual SINR constraints at different IRs, given by  $\gamma_i, i \in \mathcal{K}_{\mathcal{I}}$ . Denote  $\alpha_j$  as the given energy weight for ER  $j$ ,  $\alpha_j \geq 0$ , where larger weight value of  $\alpha_j$  indicates higher priority of transferring energy to ER  $j$  as compared to other ERs. Define  $\mathbf{G} = \zeta \sum_{j \in \mathcal{K}_{\mathcal{E}}} \alpha_j \mathbf{h}_j^H \mathbf{h}_j$ . Then from Eq. (10.5) the weighted sum-power harvested by all ERs can be expressed as  $\sum_{j \in \mathcal{K}_{\mathcal{E}}} \alpha_j Q_j = \sum_{i \in \mathcal{K}_{\mathcal{I}}} \mathbf{v}_i^H \mathbf{G} \mathbf{v}_i + \sum_{j \in \mathcal{K}_{\mathcal{E}}} \mathbf{w}_j^H \mathbf{G} \mathbf{w}_j$ . The design problems by assuming that all IRs are of either Type I or Type II are thus formulated accordingly as follows, respectively.

$$\begin{aligned} (\text{P1}) : \max_{\{\mathbf{v}_i\}, \{\mathbf{w}_j\}} \quad & \sum_{i \in \mathcal{K}_{\mathcal{I}}} \mathbf{v}_i^H \mathbf{G} \mathbf{v}_i + \sum_{j \in \mathcal{K}_{\mathcal{E}}} \mathbf{w}_j^H \mathbf{G} \mathbf{w}_j \\ \text{s.t.} \quad & \text{SINR}_i^{(\text{I})} \geq \gamma_i, \quad \forall i \in \mathcal{K}_{\mathcal{I}} \\ & \sum_{i \in \mathcal{K}_{\mathcal{I}}} \|\mathbf{v}_i\|^2 + \sum_{j \in \mathcal{K}_{\mathcal{E}}} \|\mathbf{w}_j\|^2 \leq P. \\ (\text{P2}) : \max_{\{\mathbf{v}_i\}, \{\mathbf{w}_j\}} \quad & \sum_{i \in \mathcal{K}_{\mathcal{I}}} \mathbf{v}_i^H \mathbf{G} \mathbf{v}_i + \sum_{j \in \mathcal{K}_{\mathcal{E}}} \mathbf{w}_j^H \mathbf{G} \mathbf{w}_j \\ \text{s.t.} \quad & \text{SINR}_i^{(\text{II})} \geq \gamma_i, \quad \forall i \in \mathcal{K}_{\mathcal{I}} \\ & \sum_{i \in \mathcal{K}_{\mathcal{I}}} \|\mathbf{v}_i\|^2 + \sum_{j \in \mathcal{K}_{\mathcal{E}}} \|\mathbf{w}_j\|^2 \leq P. \end{aligned}$$

Notice that the only difference between (P1) and (P2) lies in the achievable SINR expression for each IR  $i \in \mathcal{K}_{\mathcal{I}}$ . Both problems (P1) and (P2) can be shown to maximize a convex quadratic function with  $\mathbf{G}$  being positive semidefinite, i.e.,  $\mathbf{G} \succeq \mathbf{0}$ , subject to various quadratic constraints; thus they are both nonconvex quadratically constrained quadratic programs (QCQPs) [16], for which the globally optimal solutions are difficult to obtain in general. Prior to solving these two problems, we first have a check on their feasibility, i.e., whether a given set of SINR constraints for IRs can be met under the given transmit sum-power constraint  $P$ . It can be observed from (P1) and (P2) that both problems are feasible if and only if their feasibility is guaranteed by ignoring all the ERs, i.e., setting  $\alpha_j = 0$  and  $\mathbf{w}_j = \mathbf{0}, \forall j \in \mathcal{K}_{\mathcal{E}}$ . It then follows that  $\text{SINR}_i^{(\text{I})} = \text{SINR}_i^{(\text{II})}, \forall i \in \mathcal{K}_{\mathcal{I}}$ . For convenience, we denote  $\text{SINR}_i^{(\text{I})} = \text{SINR}_i^{(\text{II})} \triangleq \text{SINR}_i, \forall i \in \mathcal{K}_{\mathcal{I}}$  in this case. Thus, the feasibility of both (P1) and (P2) can be verified by solving the following problem:

$$\begin{aligned} \text{find} \quad & \{\mathbf{v}_i\} \\ \text{s.t.} \quad & \text{SINR}_i \geq \gamma_i, \quad \forall i \in \mathcal{K}_{\mathcal{I}} \\ & \sum_{i \in \mathcal{K}_{\mathcal{I}}} \|\mathbf{v}_i\|^2 \leq P. \end{aligned} \quad (10.6)$$

Problem (10.6) can be solved by the standard interior point method via transforming it into an SOCP [8] or by an uplink-downlink duality-based fixed-point iteration algorithm [10].

Next, we consider the other extreme case with no IRs, i.e.,  $\mathcal{K}_{\mathcal{I}} = \emptyset$ , where by setting  $\mathbf{v}_i = 0, \gamma_i = 0, \forall i \in \mathcal{K}_{\mathcal{I}}$ , both (P1) and (P2) are reduced to

$$\begin{aligned} & \max_{\{\mathbf{w}_j\}} \sum_{j \in \mathcal{K}_{\mathcal{E}}} \mathbf{w}_j^H \mathbf{G} \mathbf{w}_j \\ & \text{s.t. } \sum_{j \in \mathcal{K}_{\mathcal{E}}} \|\mathbf{w}_j\|^2 \leq P. \end{aligned} \quad (10.7)$$

Let  $\xi_E$  and  $\mathbf{w}_E$  be the dominant eigenvalue and its corresponding eigenvector of  $\mathbf{G}$ , respectively. Then it can be easily shown that the optimal value of Eq. (10.7) is  $\xi_E P$ , which is attained by setting  $\mathbf{w}_j = \sqrt{q_j} \mathbf{w}_E, \forall j \in \mathcal{K}_{\mathcal{E}}$ , for any set of  $q_j \geq 0, \forall j \in \mathcal{K}_{\mathcal{E}}$  satisfying  $\sum_{j \in \mathcal{K}_{\mathcal{E}}} q_j = P$ . Accordingly, all energy beams are aligned to the same direction as  $\mathbf{w}_E$ . Thus, without loss of optimality, we can set  $\mathbf{w}_j = \sqrt{P} \mathbf{w}_E$  for any  $j \in \mathcal{K}_{\mathcal{E}}$  and  $\mathbf{w}_k = \mathbf{0}, \forall k \in \mathcal{K}_{\mathcal{E}}, k \neq j$ . For convenience, we refer to the beamformer in the form of  $\sqrt{P} \mathbf{w}_E$  as the optimal energy beamformer (OeBF).

Finally, we consider another special case with all  $\gamma_i$ 's being sufficiently small (but still nonzero in general), namely the “OeBF-feasible” case, in which aligning all information beams to the OeBF is feasible for both (P1) and (P2). In other words, there exists a solution to the feasibility problem (10.6) given by  $\mathbf{v}_i = \sqrt{p_i} \mathbf{w}_E, \forall i \in \mathcal{K}_{\mathcal{I}}$  with  $p_i \geq 0, \forall i \in \mathcal{K}_{\mathcal{I}}$  satisfying  $\sum_{i \in \mathcal{K}_{\mathcal{I}}} p_i \leq P$ , i.e., the following problem has a feasible solution given by  $\{p_i\}$ .

$$\begin{aligned} & \text{find } \{p_i\} \\ & \text{s.t. } \frac{p_i |\mathbf{h}_i \mathbf{w}_E|^2}{\sum_{k \neq i, k \in \mathcal{K}_{\mathcal{I}}} p_k |\mathbf{h}_k \mathbf{w}_E|^2 + \sigma_i^2} \geq \gamma_i, \forall i \in \mathcal{K}_{\mathcal{I}} \\ & \quad \sum_{i \in \mathcal{K}_{\mathcal{I}}} p_i \leq P. \end{aligned} \quad (10.8)$$

In this case, it is easy to verify that the optimal values of both problems (P1) and (P2) are  $\xi_E P$ , which is the same as that of problem (10.7) and can be attained by

$$\mathbf{v}_i = \sqrt{\frac{P}{\sum_{k \in \mathcal{K}_{\mathcal{I}}} p_k}} \mathbf{p}_i \mathbf{w}_E, \quad \forall i \in \mathcal{K}_{\mathcal{I}} \text{ satisfying } \sum_{i \in \mathcal{K}_{\mathcal{I}}} \|\mathbf{v}_i\|^2 = P, \text{ and } \mathbf{w}_j = \mathbf{0}, \forall j \in \mathcal{K}_{\mathcal{E}},$$

i.e., no dedicated energy beam is needed to achieve the maximum weighted sum-power for ERs. To check whether the OeBF-feasible case occurs or not, we only need to solve the feasibility problem in Eq. (10.8) which is a simple linear program (LP). Therefore, in the rest of this paper, we will mainly focus on the unaddressed nontrivial case so far when (P1) and (P2) are both feasible but aligning all information beams to the OeBF is infeasible for both problems, unless otherwise specified.

### 10.2.1.3 Optimal solution via SDR

Next, we study the two nonconvex QCQPs in (P1) and (P2), and derive their optimal solutions via semidefinite relaxation (SDR). For nonconvex QCQPs, it is known that SDR is an efficient approach to obtain good approximate solutions in general [17].

In the following, by applying SDR and exploiting the specific problem structures, the globally optimal solutions for both (P1) and (P2) are obtained efficiently. Please refer to [13] for the detailed proofs about the propositions presented in this subsection.

First, consider problem (P1) for the case of Type I IRs. Define the following matrices:  $\mathbf{V}_i = \mathbf{v}_i \mathbf{v}_i^H, \forall i \in \mathcal{K}_{\mathcal{I}}$  and  $\mathbf{V}_E = \sum_{j \in \mathcal{K}_{\mathcal{E}}} \mathbf{w}_j \mathbf{w}_j^H$ . Then, it follows that  $\text{rank}(\mathbf{V}_i) \leq 1, \forall i \in \mathcal{K}_{\mathcal{I}}$  and  $\text{rank}(\mathbf{V}_E) \leq \min(M, K_E)$ . By ignoring the above rank constraints on  $\mathbf{V}_i$ 's and  $\mathbf{V}_E$ , the SDR of (P1) is given by

$$\begin{aligned} (\text{P1-SDR}): \quad & \max_{\{\mathbf{V}_i\}, \mathbf{V}_E} \sum_{i \in \mathcal{K}_{\mathcal{I}}} \text{tr}(\mathbf{G}\mathbf{V}_i) + \text{tr}(\mathbf{G}\mathbf{V}_E) \\ \text{s.t.} \quad & \frac{\text{tr}(\mathbf{h}_i^H \mathbf{h}_i \mathbf{V}_i)}{\gamma_i} - \sum_{k \neq i, k \in \mathcal{K}_{\mathcal{I}}} \text{tr}(\mathbf{h}_i^H \mathbf{h}_i \mathbf{V}_k) - \text{tr}(\mathbf{h}_i^H \mathbf{h}_i \mathbf{V}_E) - \sigma_i^2 \geq 0, \quad \forall i \in \mathcal{K}_{\mathcal{I}} \\ & \sum_{i \in \mathcal{K}_{\mathcal{I}}} \text{tr}(\mathbf{V}_i) + \text{tr}(\mathbf{V}_E) \leq P \\ & \mathbf{V}_i \succeq \mathbf{0}, \quad \forall i \in \mathcal{K}_{\mathcal{I}}, \quad \mathbf{V}_E \succeq \mathbf{0}. \end{aligned}$$

Let the optimal solution of (P1-SDR) be  $\mathbf{V}_i^*, \forall i \in \mathcal{K}_{\mathcal{I}}$  and  $\mathbf{V}_E^*$ . Then we have the following proposition.

**Proposition 10.1** *For the case of Type I IRs, the optimal solution of (P1-SDR) satisfies:  $\text{rank}(\mathbf{V}_i^*) = 1, \forall i \in \mathcal{K}_{\mathcal{I}}$ , and  $\mathbf{V}_E^* = \mathbf{0}$ .*

From Proposition 10.1, it follows that the optimal solution of (P1-SDR) satisfies the desired rank constraints, and thus the globally optimal solution of (P1) can always be obtained by solving (P1-SDR). Note that (P1-SDR) is a semidefinite program (SDP), which can be efficiently solved by existing software, e.g., CVX. Furthermore, it is observed that the optimal solution satisfies that  $\mathbf{V}_E^* = \mathbf{0}$  for (P1-SDR) or equivalently  $\mathbf{w}_j = \mathbf{0}, \forall j \in \mathcal{K}_{\mathcal{E}}$  for (P1), which implies that no dedicated energy beam is needed for achieving the maximum weighted sum harvested power in (P1). This can be intuitively explained as follows. Since Type I IRs cannot cancel the interference from energy beams (if any), employing energy beams will increase the interference power and as a result degrade the SINR at IRs. Thus, the optimal transmission strategy is to adjust the weights and power allocation of information beams only to maximize the weighted sum-power transferred to ERs.

Next, consider problem (P2) for the case of Type II IRs. Similar to (P1), the SDR of (P2) can be expressed as

$$\begin{aligned} (\text{P2-SDR}): \quad & \max_{\{\mathbf{V}_i\}, \mathbf{V}_E} \sum_{i \in \mathcal{K}_{\mathcal{I}}} \text{tr}(\mathbf{G}\mathbf{V}_i) + \text{tr}(\mathbf{G}\mathbf{V}_E) \\ \text{s.t.} \quad & \frac{\text{tr}(\mathbf{h}_i^H \mathbf{h}_i \mathbf{V}_i)}{\gamma_i} - \sum_{k \neq i, k \in \mathcal{K}_{\mathcal{I}}} \text{tr}(\mathbf{h}_i^H \mathbf{h}_i \mathbf{V}_k) - \sigma_i^2 \geq 0, \quad \forall i \in \mathcal{K}_{\mathcal{I}} \\ & \sum_{i \in \mathcal{K}_{\mathcal{I}}} \text{tr}(\mathbf{V}_i) + \text{tr}(\mathbf{V}_E) \leq P \\ & \mathbf{V}_i \succeq \mathbf{0}, \quad \forall i \in \mathcal{K}_{\mathcal{I}}, \quad \mathbf{V}_E \succeq \mathbf{0}. \end{aligned}$$

Let the optimal solution of (P2-SDR) be  $\mathbf{V}_i^*$ ,  $\forall i \in \mathcal{K}_{\mathcal{I}}$  and  $\mathbf{V}_E^*$ . We then have the following proposition.

**Proposition 10.2** *For the case of Type II IRs, the optimal solution of (P2-SDR) satisfies:  $\text{rank}(\mathbf{V}_i^*) = 1$ ,  $\forall i \in \mathcal{K}_{\mathcal{I}}$ ,  $\text{rank}(\mathbf{V}_E^*) \leq 1$ ; furthermore, it holds that  $\mathbf{V}_E^* = q^* \mathbf{w}_E \mathbf{w}_E^H$  with  $0 \leq q^* \leq P$ .*

Based on [Proposition 10.2](#), we can obtain the globally optimal solution of (P2) by solving (P2-SDR) via CVX. Meanwhile, since  $\mathbf{V}_E^* = q^* \mathbf{w}_E \mathbf{w}_E^H$ , all energy beams should be aligned to  $\mathbf{w}_E$ , the same direction as the OeBF. Similar to problem [\(10.7\)](#), in this case, we can choose to send only one energy beam to minimize the complexity of beamforming implementation at the transmitter as well as the energy signal interference cancelation at all IRs by setting  $\mathbf{w}_j = \sqrt{q^*} \mathbf{w}_E$  for any  $j \in \mathcal{K}_{\mathcal{E}}$  and  $\mathbf{w}_k = \mathbf{0}$ ,  $\forall k \in \mathcal{K}_{\mathcal{E}}, k \neq j$ .

By comparing the optimal solutions for (P1) and (P2), we can see that their main difference lies in that whether or not energy beamforming is employed. Note that the optimal value of (P2) is in general an upper bound on that of (P1) since any feasible solution of (P1) is also feasible for (P2), but not vice versa in general. If  $q^* = 0$  in [Proposition 10.2](#), then the upper bound is tight; however, if  $q^* > 0$ , then a higher weighted sum harvested power is achievable for ERs with Type II IRs. Therefore, the benefit of using Type II IRs can be realized by employing no more than one energy beam and at the cost of implementing an additional interference cancelation (with a priori known energy signals) at IRs. Nevertheless, it is also worth pointing out an interesting case with one single IR, for which energy beamforming is always not needed, as stated in the following proposition.

**Proposition 10.3** *For the case of Type II IRs, if  $K_{\mathcal{I}} = 1$ , then the optimal solution of (P2-SDR) satisfies that  $\mathbf{V}_E^* = \mathbf{0}$ .*

*Remark 10.1.* It is interesting to point out that in addition to SDR, the uplink-downlink duality has been widely used as another efficient tool to solve nonconvex transmit beamforming optimization problems in MISO/MIMO broadcast channel in wireless communication, e.g., SINR balancing in [\[10\]](#), transmit power minimization in [\[8, 18\]](#), and capacity region computation in [\[19\]](#). Interestingly, it has been shown in [\[13\]](#) that the uplink-downlink duality can also be utilized to solve (P1) and (P2) here. In particular, by leveraging the fact that the SDRs are tight for both QCQPs, we can reformulate the QCQP problem for each IR type to an equivalent transmit power minimization problem for the MISO broadcast channel with information transmission only, based upon which a new form of “uplink-downlink” duality is established for SWIPT.

#### 10.2.1.4 Numerical examples

We provide numerical examples to validate our results. We assume that the signal attenuation from the AP to all ERs is 30 dB corresponding to an equal distance of 1 m, and that to all IRs is 70 dB at an equal distance of 20 m. The channel vector  $\mathbf{h}_k$ 's are randomly generated from i.i.d. Rayleigh fading with the average channel

powers set according to the above average attenuation values. We set  $P = 1$  Watt (W) or 30 dBm,  $\zeta = 50\%$ ,  $\sigma_i^2 = -50$  dBm, and  $\gamma_i = \gamma, \forall i \in \mathcal{K}_T$ . We also set  $\alpha_j = \frac{1}{K_E}, \forall j \in \mathcal{K}_E$ ; thus the average harvested power of all ERs is considered.

Fig. 10.4 compares the average harvested power obtained by solving (P1) for Type I IRs and that by (P2) for Type II IRs versus different SINR constraint values of  $\gamma$  with fixed  $M = 4$  and  $K_E = 2$  and over 200 random channel realizations. It is observed that Type I and Type II IRs have the same performance when  $K_I = 1$ , which is consistent with Proposition 10.3. With  $K_I = 2$  or 4, it is observed that Type I and Type II IRs have similar performance when  $\gamma$  is either large or small, while the latter outperforms the former notably for moderate values of  $\gamma$ . The reasons can be explained as follows. When  $\gamma$  is sufficiently small, aligning all information beams in the direction of the OeBF is not only feasible but also optimal for both (P1) and (P2); thus, the same performance for both types of IRs is observed in Fig. 10.4. On the other hand, when  $\gamma$  is sufficiently large, it is optimal to allocate all transmit power to information beams to ensure that the SINR constraints at IRs are all met; as a result, transmit power allocated to energy beams is zero for both types of IRs, and thus their performances are also identical. At last, for the case of intermediate values of  $\gamma$ , the considerable performance gain by Type II over Type I IRs is due to the use of one dedicated energy beam. For

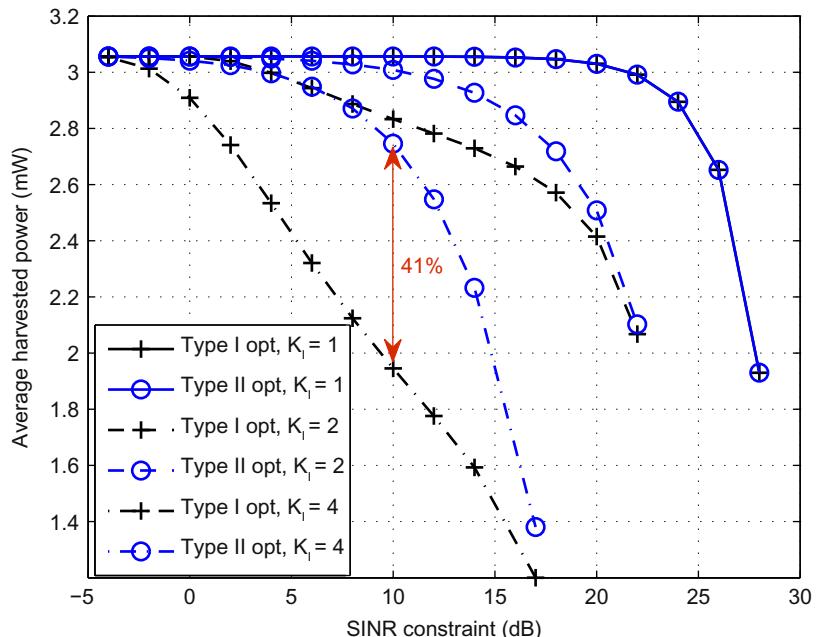


FIG. 10.4

Average harvested power versus SINR constraint with optimal beamforming designs.

example, as shown in Fig. 10.4, a 41% average harvested power gain is achieved for ERs with Type II IRs as compared to Type I IRs when  $\gamma = 10$  dB and  $K_I = 4$ , thanks to the cancelation of (known) energy signals at IRs.

### 10.2.2 SECRECY BEAMFORMING DESIGN FOR SWIPT

In the previous subsection, a receiver location based scheduling has been considered in the SWIPT system with separate IRs and ERs, where the receivers in more proximity to the AP are scheduled for WPT, while the others are scheduled for WIT. With this scheduling scheme, the optimal beamforming design is obtained to maximize the weighted sum-energy harvested by all the ERs subject to the SINR constraints of the IRs. However, although the information messages are conveyed to the IRs with sufficiently high SINRs by applying the proposed beamforming strategy, they are susceptible to the interception of the ERs, which are closer to the AP than the IRs, and thus receive IRs' information signals with better channels. To solve this potential security issue for IRs in SWIPT, in this subsection we study joint information and energy beamforming design that achieves the following two goals at the same time: secure WIT to the IRs and efficient WPT to the ERs.

According to Proposition 10.1 in Section 10.2.1 (for the Type I receivers without the cancelation of energy signal), without taking the information security into account, the optimal transmit beamforming strategy for SWIPT is to allocate all the transmit power to the information beams such that the IRs and ERs can decode the information and harvest the energy from them, respectively. However, this beamforming strategy is in general infeasible in a secure SWIPT system explained as follows. For the secure information transmission to the IRs, the aforementioned strategy requires the received information signals at the ERs to be sufficiently weak to prevent the ERs' eavesdropping, which however severely limits the amount of harvested energy by ERs. To resolve this conflict, unlike in Section 10.2.1, the transmit signal at the AP for secure SWIPT should constitute both the information signals for the IRs and the energy signals for the ERs. With a careful design of the information and energy beamforming vectors, the information signals are strongly received only at the intended IRs, while the energy signals have high power only at the ERs such that they can provide sufficient energy to the ERs and also play the role of AN to decrease the eavesdropping SINRs of the ERs to achieve both efficient and secure WPT. The details are given as follows.

#### 10.2.2.1 System model

For the purpose of exposition, we assume that in the considered secure SWIPT system, there is only one IR, which is denoted by IR 1, while the remaining  $K - 1$  users are all ERs, which are denoted by ERs 2, ...,  $K$ , i.e.,  $\mathcal{K}_I = \{1\}$  and  $\mathcal{K}_E = \{2, \dots, K\}$ . The other setups of the system are the same as in Section 10.2.1. We assume that the transmit signal consists of both the information and energy beams as explained above, where the IR is assigned with one dedicated information beam, while the

$K - 1$  ERs are in total assigned with  $d \leq M$  energy beams without loss of generality. Therefore, the complex baseband transmitted signal of the AP can be expressed as

$$\mathbf{x} = \mathbf{v}_1 s_1^{\text{ID}} + \sum_{i=1}^d \mathbf{w}_i s_i^{\text{EH}}, \quad (10.9)$$

where  $\mathbf{v}_1 \in \mathbb{C}^{M \times 1}$  and  $\mathbf{w}_i \in \mathbb{C}^{M \times 1}$  denote the information beamforming vector and the  $i$ th energy beamforming vector,  $1 \leq i \leq d$ , respectively;  $s_1^{\text{ID}}$  denotes the transmitted signal for the IR, while  $s_i^{\text{EH}}$ 's,  $i = 1, \dots, d$ , denote the energy-carrying signals for energy beams. It is assumed that  $s_1^{\text{ID}}$  is a CSCG random variable with zero mean and unit variance, denoted by  $s_1 \sim \mathcal{CN}(0, 1)$ . Furthermore,  $s_i^{\text{EH}}$ ,  $1 \leq i \leq d$ , in general can be arbitrary independent random signals each with zero-mean and unit average power. Since in this subsection we consider secret information transmission to the IR, the energy signals  $s_i^{\text{EH}}$ ,  $1 \leq i \leq d$ , also play the role of AN to reduce the information rate eavesdropped by ERs [20]. As a result, similarly as Ref. [20], we assume that  $s_i^{\text{EH}}$ 's are i.i.d. CSCG random variables denoted by  $s_i^{\text{EH}} \sim \mathcal{CN}(0, 1)$ ,  $\forall i$ , since the worst-case noise distribution for the eavesdropping ERs is known to be Gaussian. Suppose that Tx has a transmit sum-power constraint  $\bar{P}$ ; from Eq. (10.9), we thus have  $E[\mathbf{x}^H \mathbf{x}] = \|\mathbf{v}_1\|^2 + \sum_{i=1}^d \|\mathbf{w}_i\|^2 \leq \bar{P}$ .

With the transmitted signal given in Eq. (10.9), the signal received at user  $k$  is expressed as

$$y_k = \mathbf{h}_k^H \mathbf{v}_1 s_1^{\text{ID}} + \mathbf{h}_k^H \sum_{i=1}^d \mathbf{w}_i s_i^{\text{EH}} + z_k, \quad k = 1, \dots, K, \quad (10.10)$$

where  $z_k \sim \mathcal{CN}(0, \sigma_k^2)$  denotes the AWGN at user  $k$ . It is assumed that  $z_k$ 's are independent over  $k$ .

According to Eq. (10.10), the SINR at the IR can be expressed as

$$\gamma_1 = \frac{|\mathbf{v}_1^H \mathbf{h}_1|^2}{\sum_{i=1}^d |\mathbf{w}_i^H \mathbf{h}_1|^2 + \sigma_1^2}. \quad (10.11)$$

Furthermore, the SINR at  $\text{ER}_k$  (suppose that it is an eavesdropper to decode the message for the IR instead of harvesting energy only) can be expressed as

$$\gamma_k = \frac{|\mathbf{v}_1^H \mathbf{h}_k|^2}{\sum_{i=1}^d |\mathbf{w}_i^H \mathbf{h}_k|^2 + \sigma_k^2}, \quad k = 2, \dots, K. \quad (10.12)$$

The achievable secrecy rate at IR is thus given by Liang et al. [21]:

$$r_1 = \min_{2 \leq k \leq K} \log_2(1 + \gamma_1) - \log_2(1 + \gamma_k). \quad (10.13)$$

Notice that the above achievable rate may be a conservative one in practical SWIPT systems since it is unlikely that all ERs will not harvest energy but instead eavesdrop information for the IR.

On the other hand, for WPT, due to the broadcast nature of wireless channels, the energy carried by all information and energy beams, i.e.,  $\mathbf{v}_1$  and  $\mathbf{w}_i$ 's ( $1 \leq i \leq d$ ), can all be harvested at each ER. Hence, assuming unit slot duration, the harvested energy of ER<sub>k</sub> in each slot is given by Zhang and Ho [4]:

$$E_k = \zeta \left( |\mathbf{v}_1^H \mathbf{h}_k|^2 + \sum_{i=1}^d |\mathbf{w}_i^H \mathbf{h}_k|^2 \right), \quad 2 \leq k \leq K. \quad (10.14)$$

### 10.2.2.2 Problem formulation

Similar to Section 10.2.1, we are interested in maximizing the weighted sum-energy harvested by all the ERs, but subject to the secrecy rate constraint of the IR. Specifically, the problem is formulated as

$$\begin{aligned} (\text{P3}): \quad & \underset{\mathbf{v}_1, \{\mathbf{w}_i\}}{\text{Maximize}} \quad \sum_{k=2}^K \mu_k \zeta \left( |\mathbf{v}_1^H \mathbf{h}_k|^2 + \sum_{i=1}^d |\mathbf{w}_i^H \mathbf{h}_k|^2 \right) \\ & \text{Subject to} \quad r_1 \geq \bar{r}_1, \\ & \quad \|\mathbf{v}_1\|^2 + \sum_{i=1}^d \|\mathbf{w}_i\|^2 \leq \bar{P}, \end{aligned}$$

where  $\mu_k \geq 0$  denotes the energy weight for ER<sub>k</sub>, and  $\bar{r}_1$  is the target secrecy rate for IR.

It is observed that instead of an SINR constraint of the IR as in problems (P1) and (P2), problem (P3) considers a secrecy rate constraint for the IR. Due to this difference, the optimality of one single information beam shown in Proposition 10.1 does not hold for problem (P3) in general. Specifically, the information beam alone cannot achieve the secure information transmission to the IR, which requires  $|\mathbf{v}_1^H \mathbf{h}_k|^2$  to be small at each ER to avoid any “leakage” information, while efficient power transfer to the ERs requires  $|\mathbf{v}_1^H \mathbf{h}_k|^2$  to be as large as possible at each ER. Thus, the dual use of the energy beams is the key to resolve the above conflict, i.e., we need to properly design the energy beamforming vectors  $\mathbf{w}_i$ ,  $i = 1, \dots, d$ , to not only provide sufficient wireless energy transmitted to ERs, but also play the important role of AN to reduce the ERs' SINR for decoding the IR's message.

It is also observed that problem (P3) is nonconvex in general, since both the secrecy rate  $r_1$  for IR given in Eq. (10.13) and the harvested energy  $E_k$  of ER<sub>k</sub> given in Eq. (10.14) are nonconcave functions with respect to  $\mathbf{v}_0$  and  $\mathbf{w}_i$ 's. As a result, the conventional convex optimization technique cannot be directly used to solve problem (P3) globally.

### 10.2.2.3 Optimal beamforming solution

In this subsection, we propose a SDR-based algorithm to solve problem (P3) optimally by reformulating it into two subproblems. Specifically, it can be shown that there always exists an SINR constraint  $\bar{\gamma}_1 > 0$  at the IR such that the following problem,

$$\begin{aligned}
(P3.1) : \quad & \underset{\mathbf{v}_1, \{\mathbf{w}_i\}}{\text{Maximize}} && \sum_{k=2}^K \mu_k \zeta \left( |\mathbf{v}_1^H \mathbf{h}_k|^2 + \sum_{i=1}^d |\mathbf{w}_i^H \mathbf{h}_k|^2 \right) \\
& \text{Subject to} && \frac{|\mathbf{v}_1^H \mathbf{h}_1|^2}{\sum_{i=1}^d |\mathbf{w}_i^H \mathbf{h}_1|^2 + \sigma_1^2} \geq \bar{\gamma}_1, \\
& && \frac{|\mathbf{v}_1^H \mathbf{h}_k|^2}{\sum_{i=1}^d |\mathbf{w}_i^H \mathbf{h}_k|^2 + \sigma_k^2} \leq \frac{1 + \bar{\gamma}_1}{2^{\bar{\gamma}_1}} - 1, \quad k = 2, \dots, K, \\
& && \|\mathbf{v}_1\|^2 + \sum_{i=1}^d \|\mathbf{w}_i\|^2 \leq \bar{P},
\end{aligned}$$

has the same optimal solution to (P3) [22]. Furthermore, let  $g(\bar{\gamma}_1)$  denote the optimal value of problem (P3.1) with a given  $\bar{\gamma}_1 > 0$ , then the optimal value of problem (P3) is the same as that of the following problem [22]

$$(P3.2) : \underset{\bar{\gamma}_1 > 0}{\text{Maximize}} \quad g(\bar{\gamma}_1).$$

Let  $\bar{\gamma}_1^*$  denote the optimal solution to problem (P3.2). From the above results, with  $\bar{\gamma}_1 = \bar{\gamma}_1^*$ , it follows that problems (P3) and (P3.1) have the same optimal solution. Therefore, problem (P3) can be solved in the following two steps: First, given any  $\bar{\gamma}_1 > 0$ , we solve problem (P3.1) to find  $g(\bar{\gamma}_1)$ ; then, we solve problem (P3.2) to obtain the optimal  $\bar{\gamma}_1^*$  by a one-dimensional search over  $\bar{\gamma}_1 > 0$ . In the following, we focus on solving (P3.1), which is nonconvex.

Define  $\mathbf{S} = \mathbf{v}_1 \mathbf{v}_1^H$  and  $\mathbf{Q} = \sum_{i=1}^d \mathbf{w}_i \mathbf{w}_i^H$ . Then, by ignoring the rank-one constraints of  $\mathbf{S}$  and  $\mathbf{Q}$ , the SDR of (P3.1) can be expressed as

$$\begin{aligned}
(P3.1 - \text{SDR}) : \quad & \underset{\mathbf{S}, \mathbf{Q}}{\text{Maximize}} && \sum_{k=2}^K \mu_k \zeta (\text{Tr}(\mathbf{H}_1 \mathbf{S}) + \text{Tr}(\mathbf{H}_k \mathbf{Q})) \\
& \text{Subject to} && \text{Tr}(\mathbf{H}_1 \mathbf{S}) \geq \bar{\gamma}_1 (\text{Tr}(\mathbf{H}_1 \mathbf{Q}) + \sigma_1^2), \\
& && \frac{\text{Tr}(\mathbf{H}_k \mathbf{S})}{\bar{\gamma}_e} \leq \text{Tr}(\mathbf{H}_k \mathbf{Q}) + \sigma_k^2, \quad \forall k, \\
& && \text{Tr}(\mathbf{S}) + \text{Tr}(\mathbf{Q}) \leq \bar{P}, \\
& && \mathbf{S} \succeq \mathbf{0}, \quad \mathbf{Q} \succeq \mathbf{0},
\end{aligned}$$

where  $\mathbf{H}_k = \mathbf{h}_k \mathbf{h}_k^H$ , and  $\bar{\gamma}_e = (1 + \bar{\gamma}_1)/2^{\bar{\gamma}_1} - 1$ .

Since (P3.1-SDR) is convex, it can be solved by CVX. If the optimal solution to problem (P3.1-SDR), denoted by  $\mathbf{S}^*$  and  $\mathbf{Q}^*$ , satisfies  $\text{rank}(\mathbf{S}^*) = 1$ , then the optimal information beam  $\mathbf{v}_1^*$  and energy beam  $\mathbf{w}_i^*$ 's,  $i = 1, \dots, d$  ( $d = \text{rank}(\mathbf{Q}^*)$ ), for problem (P3.1) can be obtained from the eigenvalue decompositions (EVDs) of  $\mathbf{S}^*$  and  $\mathbf{Q}^*$ , respectively; otherwise, if  $\text{rank}(\mathbf{S}^*) > 1$ , the optimal value of problem (P3.1-SDR) only serves as an upper bound on that of problem (P3.1). Fortunately, it is shown in [22] that there always exists an optimal solution  $(\mathbf{S}^*, \mathbf{Q}^*)$  to (P3.1-SDR) with  $\text{rank}(\mathbf{S}^*) = 1$ , and there is no loss of optimality for (P3.1) due to the rank relaxation on  $\mathbf{S}$  in (P3.1-SDR).

#### 10.2.2.4 Numerical results

In the following, we provide one numerical example to verify the results. In this example, we consider a MISO SWIPT system as shown in Fig. 10.5, where there are one IR and  $K = 7$  ERs. The AP is equipped with  $M = 9$  antennas. Moreover, we use the far-field uniform linear antenna array [23] to model the channels. Specifically,

$$\mathbf{h}_k = \rho_{h_k} \times [1, e^{j\theta_k}, \dots, e^{j(M-1)\theta_k}]^T, \quad k = 1, \dots, K, \quad (10.15)$$

where  $\rho_{h_1}^2 = -70$  dB (corresponding to a distance of 20 m),  $\rho_{h_2}^2 = \dots = \rho_{h_8}^2 = -30$  dB (corresponding to an equal distance of 1 m), and  $\theta_k = -\frac{2\pi d \sin(\phi_k)}{\lambda}$ ,  $k = 1, \dots, K$ , with  $d$  denoting the spacing between successive antenna elements at the AP,  $\lambda$  denoting the carrier wavelength, and  $\phi_k$  denoting the direction of user  $k$  to the AP. We set  $d = \frac{\lambda}{2}$ , and  $\{\phi_1, \dots, \phi_8\} = \{\frac{11\pi}{16}, 0, \frac{\pi}{6}, \frac{3\pi}{8}, \frac{\pi}{2}, \frac{45\pi}{64}, \frac{9\pi}{8}, \frac{13\pi}{9}\}$ . We set  $\bar{P} = 1$  Watt (W) or 30 dBm,  $\zeta = 50\%$ , and  $\sigma_k^2 = -50$  dBm,  $1 \leq k \leq K$ . We also set  $\mu_k = 1$ ,  $1 \leq k \leq K$  in (P3); thus, the sum-energy harvested by all ERs is considered.

To evaluate the performance of the proposed design compared to other benchmark schemes, in the following we introduce two suboptimal beamforming solutions to problem (P3), which can be implemented with lower complexity as compared to the proposed optimal solution.

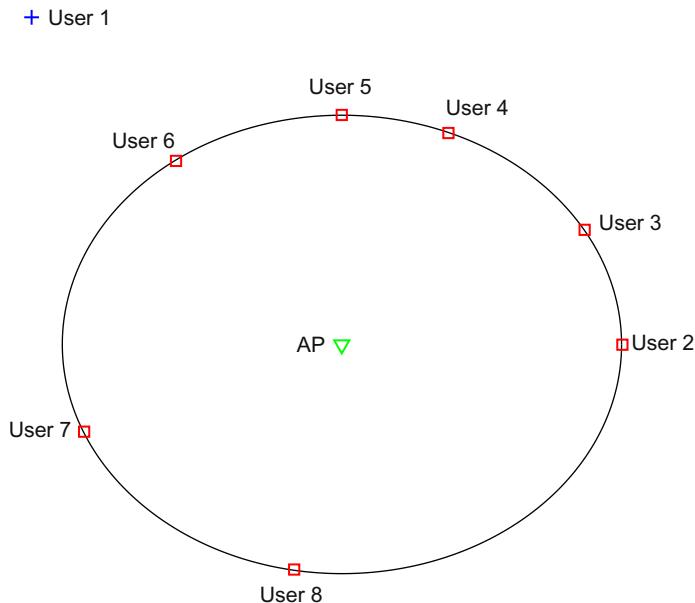


FIG. 10.5

Locations of the IR and ERs.

- **Suboptimal Solution I:** Supposing that the number of ERs is less than the number of antennas at the AP, i.e.,  $K - 1 < M$  (which is true in this numerical example since  $K = 8$  and  $M = 9$ ), then the first suboptimal solution aims to solve problem (P3) with the additional constraints:  $\mathbf{v}_1^H \mathbf{h}_k = 0$ ,  $k = 2, \dots, K$ , and  $\mathbf{w}_i^H \mathbf{h}_1 = 0$ ,  $\forall i$ . In other words, the information beam  $\mathbf{v}_1$  is aligned to the null space of the ERs' channels  $\mathbf{G} = [\mathbf{h}_2, \dots, \mathbf{h}_K]^H$  to eliminate any information leakage to the ERs, while the energy beams  $\mathbf{w}_i$ 's are all restricted to lie in the null space of the IR's channel  $\mathbf{h}_1$  such that they cause no interference to IR. The above constraints greatly reduce the complexity to solve problem (P3), and the corresponding optimal solution is given in [22] in closed-form.
- **Suboptimal Solution II:** The second suboptimal solution aims to solve problem (P3) with the additional constraints:  $\mathbf{v}_1 = \sqrt{\hat{P}_1} \mathbf{h}_1 / \|\mathbf{h}_1\|$  and  $\mathbf{w}_i^H \mathbf{h}_1 = 0$ ,  $\forall i$ , where  $\hat{P}_1 = \|\mathbf{v}_1\|^2$  denotes the transmit power of the information beam. In other words, the information beam  $\mathbf{v}_1$  is aligned to the same direction as  $\mathbf{h}_1$  to maximize the IR's SINR, while the energy beams  $\mathbf{w}_i$ 's are all restricted to lie in the null space of the IR's channel  $\mathbf{h}_1$  such that they cause no interference to IR. The optimal solution under the above constraints is also given in [22] in closed-form.

In this example, we activate one more ER at each time (from user 2 to user 8). Fig. 10.6 shows the sum-energy harvested by all the ERs by the proposed optimal

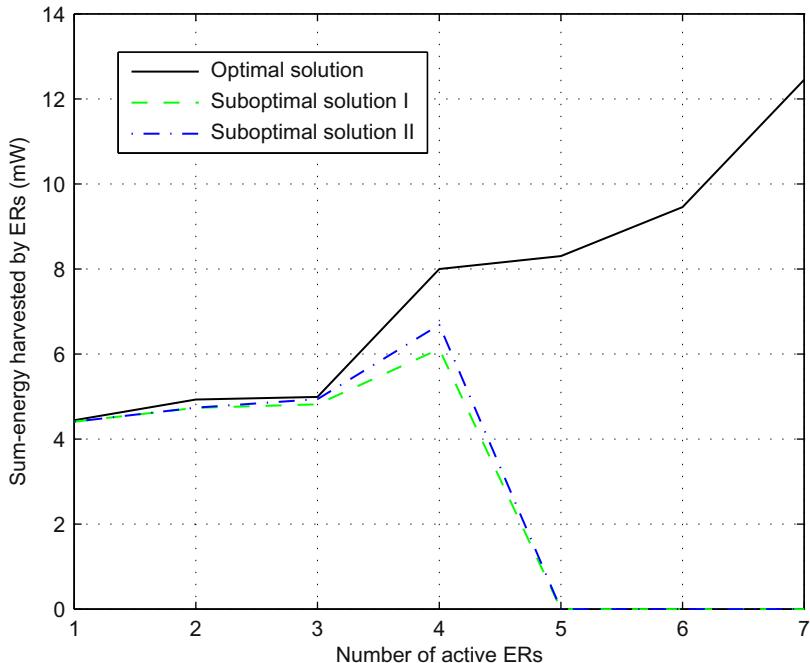


FIG. 10.6

The sum-energy harvested by ERs over the number of active ERs with given secrecy rate constraint for the IR,  $\bar{T}_0 = 4$  bps/Hz.

and suboptimal algorithms for (P3) against the number of active ERs with  $\bar{r}_1 = 4$  bps/Hz. It is observed that with more ERs, the sum-energy harvested is increased in all cases. Furthermore, it is observed that when the number of active users is no larger than 4, the performance of both Suboptimal Solutions I and II is very close to that of the optimal solution. However, after the fifth ER, i.e., user 6, is activated, both of the suboptimal solutions achieve zero sum-energy because the secrecy rate constraint cannot be satisfied in (P3). The reason is as follows. Note that for both of these two suboptimal solutions, the energy beams  $w_i$ 's are aligned into the null space of  $\mathbf{h}_1$ , i.e.,  $\mathbf{w}_i^H \mathbf{h}_1 = 0, \forall i$ . However, in this example the direction of  $\mathbf{h}_6$  is very close to that of  $\mathbf{h}_1$ . It thus follows that  $\mathbf{w}_i^H \mathbf{h}_6 \approx 0, \forall i$ . In other words, the energy beams cannot play the role of AN to reduce the SINR of user 6 in this case. However, it is observed from Fig. 10.6 that even in the challenging scenario where one ER is located in a direction very close to (but not exactly the same as) the IR, our proposed optimal algorithm still achieves good performance thanks to the jointly optimized beamforming and power allocation design.

### 10.2.3 BEAMFORMING DESIGN FOR SWIPT SYSTEM WITH CO-LOCATED IRS AND ERS

In Sections 10.2.1 and 10.2.2, we have considered the SWIPT system with separate IRs and ERs and obtained the optimal beamforming solutions for the cases with or without the information security consideration. In this subsection, we focus on the SWIPT system with co-located IR and ER users and investigate how to jointly design the transmit beamforming and the receive power splitting ratios. For simplicity, we ignore the secrecy information transmission issue in this subsection.

#### 10.2.3.1 System model

It is assumed that the linear transmit beamforming is implemented at the AP, where each user is assigned with one dedicated information beam. The complex baseband transmitted signal at the AP is thus expressed as

$$\mathbf{x} = \sum_{k=1}^K \mathbf{v}_k s_k^{\text{ID}}, \quad \forall k, \quad (10.16)$$

where  $s_k^{\text{ID}}$  denotes the transmitted data symbol for user  $k$ , and  $\mathbf{v}_k$  is the corresponding transmit beamforming vector. It is assumed that  $s_k^{\text{ID}}, k = 1, \dots, K$ , are i.i.d. CSCG random variables each with zero mean and unit variance, denoted by  $s_k^{\text{ID}} \sim \mathcal{CN}(0, 1)$ .

With the PS scheme, the received signal at each user is split to the IR and the ER by a power splitter, which divides  $\rho_k$  ( $0 \leq \rho_k \leq 1$ ) portion of the signal power to the IR, and the remaining  $1 - \rho_k$  portion of power to the ER. As a result, the signal split to the IR at user  $k$  is expressed as

$$y_k^{\text{ID}} = \sqrt{\rho_k} \mathbf{h}_k^H \sum_{i=1}^K \mathbf{v}_i s_i^{\text{ID}} + z_k, \quad (10.17)$$

where  $z_k \sim \mathcal{CN}(0, \sigma_k^2)$  denotes the AWGN for ID at user  $k$ . Accordingly, the SINR for decoding  $s_k^{\text{ID}}$  at user  $k$  is given by

$$\gamma_k = \frac{\rho_k |\mathbf{h}_k^H \mathbf{v}_k|^2}{\rho_k \sum_{j \neq k} |\mathbf{h}_k^H \mathbf{v}_j|^2 + \sigma_k^2}, \quad \forall k. \quad (10.18)$$

On the other hand, since  $1 - \rho_k$  portion of power is split to the ER, the harvested power by the ER of user  $k$  is given by

$$E_k = \zeta(1 - \rho_k) \sum_{i=1}^K |\mathbf{h}_k^H \mathbf{v}_i|^2, \quad \forall k. \quad (10.19)$$

### 10.2.3.2 Problem formulation

In this subsection, we aim to jointly design the beamforming vectors at the AP, i.e.,  $\mathbf{v}_k$ 's, and power splitting ratios at the receivers, i.e.,  $\rho_k$ 's, to minimize the total transmit power at the AP subject to each user's SINR constraint, denoted by  $\bar{\gamma}_k$ 's, as well as each user's harvested energy constraint, denoted by  $\bar{E}_k$ . This problem is formulated as follows.

$$\begin{aligned} (\text{P4}) : \quad & \underset{\mathbf{v}_k, \rho_k}{\text{Minimize}} \quad \sum_{k=1}^K \|\mathbf{v}_k\|^2 \\ & \text{Subject to} \quad \frac{\rho_k |\mathbf{h}_k^H \mathbf{v}_k|^2}{\rho_k \sum_{j \neq k} |\mathbf{h}_k^H \mathbf{v}_j|^2 + \sigma_k^2} \geq \bar{\gamma}_k, \quad \forall k, \\ & \quad \zeta(1 - \rho_k) \sum_{i=1}^K |\mathbf{h}_k^H \mathbf{v}_i|^2 \geq \bar{E}_k, \quad \forall k. \end{aligned}$$

Notice that in this paper we consider the general case that all users have nonzero SINR and harvested power targets, i.e.,  $\bar{\gamma}_k > 0$  and  $\bar{E}_k > 0, \forall k$ . As a result, the receive PS ratios at the users should satisfy  $0 < \rho_k < 1, \forall k$ .

Note that problem (P4) is nonconvex due to the coupled beamforming vectors  $\mathbf{v}_k$ 's and PS ratios  $\rho_k$ 's in both the SINR and harvested energy constraints. It is also worth noting that if we fix  $\rho_k$ 's with  $0 < \rho_k < 1, \forall k$ , the resulting beamforming optimization problem over  $\mathbf{v}_k$ 's is still nonconvex since the harvested energy at each user given in Eq. (10.19) is a convex (instead of concave) function of  $\mathbf{v}_k$ 's. Finally, notice that if we remove all the harvested power constraints and let  $\rho_k \rightarrow 1, \forall k$ , the above problem reduces to the conventional power minimization problem subject to only SINR constraints in the MISO broadcast channel, which can be efficiently solved by existing methods [10, 24, 25]. In the following subsection, we obtain the optimal solution to problem (P4). Note that, for practical implementation of all solutions, the computation takes place at the AP and then the AP sends each  $\rho_k$  to the corresponding user for implementation.

### 10.2.3.3 Optimal solution

Similar to problems (P1)–(P3), the technique of SDR plays a key role in optimally solving the nonconvex problem (P4). Specifically, define  $\mathbf{S}_k = \mathbf{v}_k \mathbf{v}_k^H$  and  $\mathbf{H}_k = \mathbf{h}_k \mathbf{h}_k^H$ ,  $\forall k$ . Then, by ignoring the rank-one constraints of  $\mathbf{S}_k$ 's, the SDR of problem (P4) is expressed as

$$\begin{aligned} (\text{P4 - SDR}) : \quad & \underset{\mathbf{S}_k, \rho_k}{\text{Minimize}} \quad \sum_{k=1}^K \text{Tr}(\mathbf{S}_k) \\ & \text{Subject to} \quad \frac{\rho_k \text{Tr}(\mathbf{H}_k \mathbf{S}_k)}{\rho_k \sum_{j \neq k} \text{Tr}(\mathbf{H}_k \mathbf{S}_j) + \sigma_k^2} \geq \bar{\gamma}_k, \quad \forall k, \\ & \quad \zeta (1 - \rho_k) \sum_{i=1}^K \text{Tr}(\mathbf{H}_k \mathbf{S}_k) \geq \bar{E}_k, \quad \forall k. \end{aligned}$$

Due to the coupling between  $\mathbf{S}_k$ 's and  $\rho_k$ 's, problem (P4-SDR) is still nonconvex. However, since  $0 < \rho_k < 1$ ,  $\forall k$ , it can be shown that problem (P4-SDR) is equivalent to the following convex problem.

$$\begin{aligned} (\text{P4 - SDR - Eqv}) : \quad & \underset{\mathbf{S}_k, \rho_k}{\text{Minimize}} \quad \sum_{k=1}^K \text{Tr}(\mathbf{S}_k) \\ & \text{Subject to} \quad \text{Tr}(\mathbf{H}_k \mathbf{S}_k) \geq \sum_{j \neq k} \text{Tr}(\mathbf{H}_k \mathbf{S}_j) \bar{\gamma}_k + \frac{\sigma_k^2 \bar{\gamma}_k}{\rho_k}, \quad \forall k, \\ & \quad \zeta \sum_{i=1}^K \text{Tr}(\mathbf{H}_k \mathbf{S}_k) \geq \frac{\bar{E}_k}{1 - \rho_k}, \quad \forall k. \end{aligned}$$

Let  $\mathbf{S}_k^*$ 's and  $\rho_k^*$ 's denote the optimal solution to problem (P4-SDR-Eqv). If  $\mathbf{S}_k^*$ 's satisfy  $\text{rank}(\mathbf{S}_k^*) = 1$ ,  $\forall k$ , then the optimal beamforming solution  $\mathbf{v}_k^*$  to problem (P4) can be obtained from the EVD of  $\mathbf{S}_k^*$ ,  $k = 1, \dots, K$ , and the optimal PS solution of problem (P4) is also given by the associated  $\rho_k^*$ 's. Otherwise, if there exists any  $k$  such that  $\text{rank}(\mathbf{S}_k^*) > 1$ , then in general the optimal value of problem (P4-SDR-Eqv) only serves as the lower bound of that of problem (P4). Fortunately, it is shown in [26] that the optimal solution to problem (P4-SDR-Eqv) must satisfy  $\text{rank}(\mathbf{S}_k^*) = 1$ ,  $\forall k$ . As a result, we can solve problem (P4) optimally based on the solution to the convex problem (P4-SDR-Eqv).

### 10.2.3.4 Numerical results

In the following, we provide one numerical example to verify our results. It is assumed that there are  $K = 4$  users and all the users have the same set of parameters, i.e.,  $\sigma_k^2 = \sigma^2$ ,  $\bar{E}_k = e$ , and  $\bar{\gamma}_k = \gamma$ ,  $\forall k$ . Moreover, the AP is assumed to be equipped with  $M = 4$  antennas. We set  $\zeta = 0.5$  and  $\sigma^2 = -70$  dBm in this example. It is further assumed that the signal attenuation from the AP to all users is 40 dB corresponding to an identical distance of 5 m. With this transmission distance, the line-of-sight (LOS) signal is dominant, and thus the Rician fading is used to model the channel. Specifically,  $\mathbf{h}_k$  is expressed as

$$\mathbf{h}_k = \sqrt{\frac{K_R}{1+K_R}} \mathbf{h}_k^{\text{LOS}} + \sqrt{\frac{1}{1+K_R}} \mathbf{h}_k^{\text{NLOS}}, \quad (10.20)$$

where  $\mathbf{h}_k^{\text{LOS}} \in \mathbb{C}^{M \times 1}$  is the LOS deterministic component,  $\mathbf{h}_k^{\text{NLOS}} \in \mathbb{C}^{M \times 1}$  denotes the Rayleigh fading component with each element being a CSCG random variable with zero mean and covariance of  $-40$  dB, and  $K_R$  is the Rician factor set to be  $5$  dB. Note that for the LOS component, we use the far-field uniform linear antenna array model [23] with  $\mathbf{h}_k$  given in Eq. (10.15). We set  $d = \frac{\lambda}{2}$ , and  $\{\phi_1, \phi_2, \phi_3, \phi_4\} = \{-30^\circ, -60^\circ, 60^\circ, 30^\circ\}$ .

To evaluate the performance of the proposed optimal solution, we compare it with two low-complexity suboptimal solutions to problem (P4) described as follows.

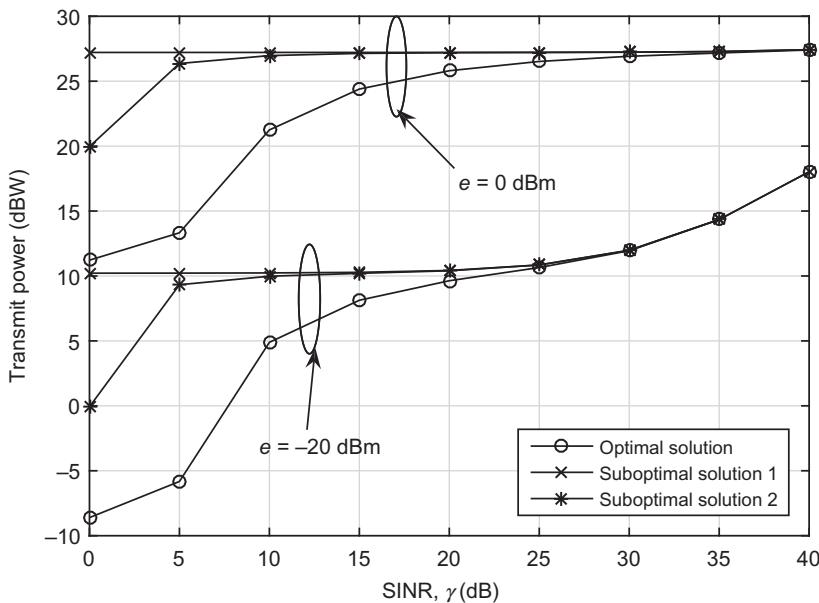
- **Suboptimal Solution 1:** Supposing that the number of users is no larger than the number of antennas at the AP, i.e.,  $K \leq M$  (which is true in this numerical example since  $K = M = 4$ ), then the first suboptimal solution aims to solve problem (P4) with the additional constraints:  $\mathbf{v}_k^H \mathbf{h}_j = 0, \forall j \neq k$ . In other words, the information beam for user  $k$ , i.e.,  $\mathbf{v}_k$ , is aligned to the null space of the channels of the other  $K - 1$  users, i.e.,  $\mathbf{G} = [\mathbf{h}_1, \dots, \mathbf{h}_{k-1}, \mathbf{h}_{k+1}, \dots, \mathbf{h}_K]^H$  to eliminate the interference. The above constraints greatly reduce the complexity to solve problem (P4), and the corresponding optimal solution can be found in [26].
- **Suboptimal Solution 2:** The second suboptimal solution aims to first solve problem (P4) without the harvested energy constraints, i.e.,

$$\begin{aligned} (\text{P4-SINR}): \quad & \underset{\mathbf{v}_k}{\text{Minimize}} \quad \sum_{k=1}^K \|\mathbf{v}_k\|^2 \\ & \text{Subject to} \quad \frac{\rho_k |\mathbf{h}_k^H \mathbf{v}_k|^2}{\rho_k \sum_{j \neq k} |\mathbf{h}_k^H \mathbf{v}_j|^2 + \sigma_k^2} \geq \bar{\gamma}_k, \quad \forall k. \end{aligned}$$

The optimal beamforming solution can be efficiently obtained based on the techniques proposed in [10, 24, 25]. Let  $\hat{\mathbf{v}}_k$ 's denote the optimal beamforming solution to problem (P4-SINR). Then, we scale up the beamformers  $\{\hat{\mathbf{v}}_k\}$  by a common factor  $\sqrt{\alpha}$  and then jointly optimize  $\alpha$  and receive PS ratios  $\rho_k$ 's to satisfy both the SINR and harvested energy constraints in problem (P4) with the minimum transmit power. The resulting problem is solved in [26].

It is shown in [26] that the above two suboptimal solutions are both asymptotically optimal when  $\gamma_k$ 's go to infinity.

First, we investigate the minimum transmit power required at the AP versus the SINR target for all users,  $\gamma$ , with their harvested power constraint,  $e$ , being fixed. Fig. 10.7 shows the performance comparison by the optimal SDR-based solution to problem (P4) and the two suboptimal solutions with  $e = 0$  dBm or  $e = -20$  dBm. It is observed that as the harvested power constraint  $e$  is increased from  $-20$  dBm to  $0$  dBm, substantially more transmit power is needed at the AP for all values of  $\gamma$ . It is also observed that for both cases of  $e = 0$  dBm and  $e = -20$  dBm, the minimum transmit power is achieved by the optimal solution for all values of  $\gamma$ . Moreover, when the SINR constraint  $\gamma$  is small, Suboptimal Solution 2 is observed to achieve notably smaller transmit power than Suboptimal Solution 1. However, as  $\gamma$  increases, the performance gap between the two suboptimal solutions vanishes. For example, when

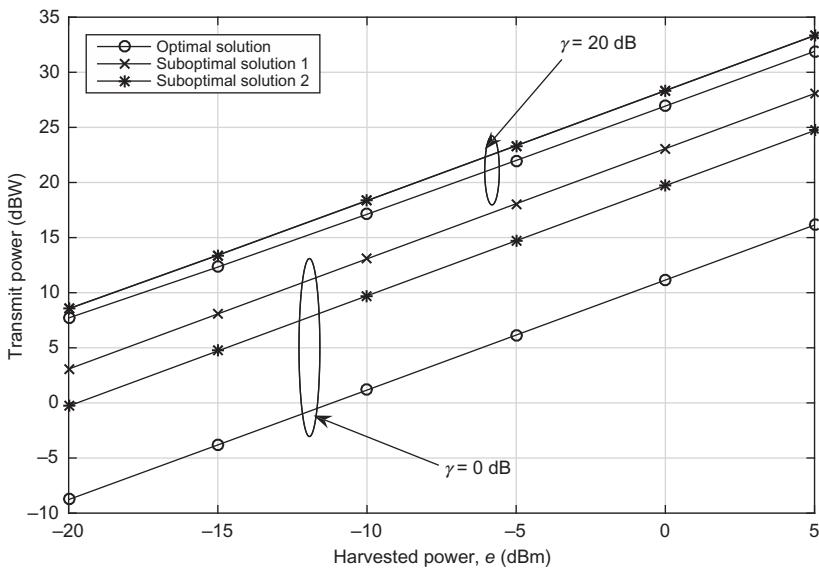
**FIG. 10.7**

Transmit power versus  $\gamma$ ,  $e = 0$  dBm or  $-20$  dBm.

$\gamma > 35$  dB for the case of  $e = 0$  dBm or  $\gamma > 25$  dB for the case of  $e = -20$  dBm, the minimum transmit power values achieved by the two suboptimal solutions both converge to that by the optimal solution.

Next, we show in Fig. 10.8 the minimum transmit power achieved by the optimal and suboptimal solutions over  $e$  with fixed  $\gamma = 0$  dB or  $\gamma = 20$  dB. Similarly as in Fig. 10.7, it is observed that for both cases of  $\gamma = 0$  dB and  $\gamma = 20$  dB, the optimal solution achieves the minimum transmit power for all values of  $e$ . Furthermore, at low SINR, i.e.,  $\gamma = 0$  dB, the transmit power achieved by Suboptimal Solution 2 is notably smaller than that by Suboptimal Solution 1, but much larger than that by the optimal solution, for both values of  $e$ . However, at high SINR, i.e.,  $\gamma \geq 20$  dB, all the optimal and suboptimal solutions perform very closely to each other. This confirms the asymptotic optimality of the suboptimal solutions as  $\gamma \rightarrow \infty$ .

*Remark 10.2.* Besides the works [13, 22, 26] introduced in the previous subsections, there are many other valuable works on the beamforming design for the MISO point-to-multipoint SWIPT systems. For example, Ng et al. [27] extends the secrecy beamforming design problem investigated in Section 10.2.2 to the case with imperfect channel state information (CSI) at the transmitter. Moreover, to reduce the complexity for the beamforming design in Section 10.2.3, Shi et al. [28] optimizes the AP's zero-forcing beamforming strategy together with each receiver's power splitting strategy for energy efficiency maximization in SWIPT systems. Interested readers may refer to [29] for more related works along this direction.

**FIG. 10.8**

Transmission power versus  $e$ ,  $\gamma = 0$  dBm or 20 dBm.

## 10.3 EXTENSIONS

So far, we have focused on the transmit beamforming design for the multiuser SWIPT system with a single transmitter and multiple receivers by assuming perfect CSI. Such designs are extendable to other system setups with various practical considerations. In the following, we discuss some extensions that have been investigated in the literature and motivate future work.

### 10.3.1 MULTIPONT-TO-MULTIPOINT SWIPT

Besides the point-to-multipoint setup, the study on the multipoint-to-multipoint SWIPT systems has been recently pursued in [30–32]. In these works, multiple transmitters send independent messages to their corresponding receivers, which at the same time also carry wireless power, and the receivers implement either TS or PS scheme to realize SWIPT. Although the multipoint-to-multipoint system can be modeled as the conventional interference channel for WIT due to the lack of joint processing and message sharing over the transmitters, the energy signal waveforms can be jointly optimized at all transmitters since unlike the information signals, the energy signals are pseudo-random and thus can be designed offline and stored for real-time transmission. Therefore, a new signal splitting scheme at the transmitters

is proposed in [33], where each transmit signal is generally split into an information signal and an energy signal for WIT and WPT, respectively, such that the energy signals of all transmitters can be jointly optimized to achieve the maximum gain of energy beamforming even though they are distributed at different locations.

### 10.3.2 WIRELESS POWERED COMMUNICATION NETWORK

Apart from the SWIPT system, another line of research related to the RF-based WIT and WPT is the so-called wireless powered communication network (WPCN), where radio signals are used to power wireless terminals for information transmission. Specifically, under the setup of one multiantenna AP and multiple single-antenna users, a harvest-then-transmit protocol is investigated in [34], where the AP first broadcasts wireless power to all users via energy beamforming in the downlink, then the users send their independent information to the AP simultaneously in the uplink using their harvested energy. With this protocol, the downlink-uplink time allocation, the downlink energy beamforming, the uplink power control and receive beamforming are jointly optimized to maximize the minimum throughput of all the users. It is shown that this nonconvex problem can be globally solved based on the nonnegative matrix theory [35]. Furthermore, a similar problem is studied in [36] where the AP is assumed to be equipped with a large number of antennas, i.e., a massive MIMO system, to improve the efficiency of WPCN. Under this setup, besides the downlink WPT and uplink WIT, the channel estimation is also considered, and the asymptotically optimal solution is obtained.

### 10.3.3 CSI ACQUISITION AT TRANSMITTER

It is worth noting that the above works have mostly assumed the perfect CSI at the AP. However, acquiring such CSI is practically challenging, especially for the ERs, since existing methods for channel learning in wireless communication [37] may not be directly applicable due to their energy and hardware limitations. To overcome such a challenge, there have been several works considering the channel acquisition methods for multiantenna WPT systems, and they can be categorized into three classes. The first class of methods exploit the channel reciprocity between the forward (from the AP to ERs) and reverse (from ERs to the AP) links [36, 38], where the AP obtains the forward link CSI by performing a reverse link channel estimation based on the training signals sent by the ER. The second class of methods send training signals directly in the forward link from the AP to the ER, through which the ER estimates the MIMO channel and then sends the estimated channel back to the AP via the reverse link [39, 40]. This method is reminiscent of the conventional channel estimation and feedback approach used in wireless communication [37]. The third class of methods are based on the energy measurement feedback by the ER [14]. In this method, the ER measures its harvested energy levels over different training intervals using an energy meter, and sends one feedback bit to the AP per interval to indicate the increase or decrease of the measured energy in the current interval as

compared to that in the previous interval. Based on the feedback bits collected in the present and past intervals, the AP adjusts its transmit beamforming in subsequent training intervals and obtains refined estimates of the MIMO channel. This energy feedback based channel learning method can be implemented without additional baseband processing modules at the ER, which are required by the first two methods.

---

## 10.4 CONCLUSION

The paper presents a new class of transmit beamforming design problems for MISO SWIPT systems under various practical setups, such as separated or co-located information and energy receivers with TS or PS, and with or without the consideration of physical-layer security. The formulated problems are shown to be nonconvex QCQPs in general, which are optimally solved by applying the techniques of SDR and exploiting the particular structure of each problem. Extensions to other setups and practical considerations are also discussed. It is our hope that this paper will serve as a useful reference for researchers to pursue their research in this exciting new direction.

---

## REFERENCES

- [1] S. Bi, C.K. Ho, R. Zhang, Wireless powered communication: opportunities and challenges, *IEEE Commun. Mag.* 53 (4) (2015) 117–125.
- [2] L.R. Varshney, Transporting information and energy simultaneously, in: Proc. IEEE Int. Symp. Inf. Theory (ISIT), 2008, pp. 1612–1616.
- [3] P. Grover, A. Sahai, Shannon meets Tesla: wireless information and power transfer, in: Proc. IEEE Int. Symp. Inf. Theory (ISIT), 2010, pp. 2363–2367.
- [4] R. Zhang, C.K. Ho, MIMO broadcasting for simultaneous wireless information and power transfer, *IEEE Trans. Wireless Commun.* 12 (5) (2013) 1989–2001.
- [5] X. Zhou, R. Zhang, C.K. Ho, Wireless information and power transfer: architecture design and rate-energy tradeoff, *IEEE Trans. Commun.* 61 (11) (2013) 4757–4767.
- [6] I.E. Telatar, Capacity of multi-antenna Gaussian channels, *Eur. Trans. Telecommun.* 10 (6) (1999) 585–596.
- [7] A.B. Gershman, N.D. Sidiropoulos, N.D.S. Shahbazpanahi, M. Bengtsson, B. Ottersten, Convex optimization based beamforming, *IEEE Sig. Proc. Mag.* 27 (3) (2010) 62–75.
- [8] A. Wiesel, Y.C. Eldar, S. Shamai, Linear precoding via conic optimization for fixed MIMO receivers, *IEEE Trans. Sig. Proc.* 54 (1) (2006) 161–176.
- [9] F. Rashid-Farrokhi, K.J.R. Liu, L. Tassiulas, Transmit beamforming and power control for cellular wireless systems, *IEEE J. Sel. Areas Commun.* 16 (8) (1998) 1437–1450.
- [10] M. Schubert, H. Boche, Solution of the multi-user downlink beamforming problem with individual SINR constraints, *IEEE Trans. Veh. Technol.* 53 (1) (2004) 18–28.
- [11] Q. Shi, M. Razaviyayn, Z. Luo, C. He, An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel, *IEEE Trans. Sig. Proc.* 9 (59) (2011) 4331–4340.

- [12] D.A. Schmidt, C. Shi, R.A. Berry, M.L. Honig, W. Utschick, Comparison of distributed beamforming algorithms for MIMO interference networks, *IEEE Trans. Sig. Proc.* 61 (13) (2013) 3476–3489.
- [13] J. Xu, L. Liu, R. Zhang, Multiuser MISO beamforming for simultaneous wireless information and power transfer, *IEEE Trans. Sig. Proc.* 62 (18) (2014) 4798–4810.
- [14] J. Xu, R. Zhang, Energy beamforming with one-bit feedback, *IEEE Trans. Sig. Proc.* 62 (20) (2014) 5370–5381.
- [15] J. Xu, R. Zhang, A general design framework for MIMO wireless energy transfer with limited feedback, *IEEE Trans. Sig. Proc.* 64 (10) (2016) 2475–2488.
- [16] S. Boyd, L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, UK, 2004.
- [17] Z.-Q. Luo, W.-K. Ma, A.M.-C. So, Y. Ye, S. Zhang, Semidefinite relaxation of quadratic optimization problems, *IEEE Sig. Proc. Mag.* 27 (3) (2010) 20–34.
- [18] W. Yu, T. Lan, Transmitter optimization for the multi-antenna downlink with per-antenna power constraints, *IEEE Trans. Sig. Proc.* 55 (6) (2007) 2646–2660.
- [19] L. Zhang, R. Zhang, Y.C. Liang, Y. Xin, H.V. Poor, On the Gaussian MIMO BC-MAC duality with multiple transmit covariance constraints, *IEEE Trans. Info. Theory* 58 (34) (2012) 2064–2078.
- [20] S. Goel, R. Negi, Guaranteeing secrecy using artificial noise, *IEEE Trans. Wireless Commun.* 7 (6) (2008) 2180–2189.
- [21] Y. Liang, G. Kramer, H.V. Poor, S. Shamai (Shitz), Compound wire-tap channels, in: *Proc. 45th Ann. Allerton Conf. Commun. Contr. Comput.* 2007, pp. 136–143.
- [22] L. Liu, R. Zhang, K.C. Chua, Secrecy wireless information and power transfer with MISO beamforming, *IEEE Trans. Sig. Proc.* 62 (7) (2014) 1850–1863.
- [23] E. Karipidis, N.D. Sidiropoulos, Z.Q. Luo, Far-field multicast beamforming for uniform linear antenna arrays, *IEEE Trans. Sig. Proc.* 55 (10) (2007) 4916–4927.
- [24] M. Bengtsson, B. Ottersten, Optimal and suboptimal transmit beamforming, in: *Handbook of Antennas in Wireless Communications*, CRC Press, Boca Raton, FL, 2001.
- [25] M. Codreanu, A. Tolli, A. Juntti, M. Latva-aho, Joint design of Tx-Rx beamformers in MIMO downlink channels, *IEEE Trans. Sig. Proc.* 55 (9) (2007) 4639–4655.
- [26] Q. Shi, L. Liu, W. Xu, R. Zhang, Joint transmit beamforming and receive power splitting for MISO SWIPT systems, *IEEE Trans. Wireless Commun.* 13 (6) (2014) 3269–3280.
- [27] D.W.K. Ng, E.S. Lo, R. Schober, Robust beamforming for secure communication in systems with wireless information and power transfer, *IEEE Trans. Wireless Commun.* 8 (13) (2014) 4599–4615.
- [28] Q. Shi, C. Peng, W. Xu, M. Hong, Y. Cai, Energy efficiency optimization for MISO SWIPT systems with zero-forcing beamforming, *IEEE Trans. Sig. Proc.* 64 (4) (2016) 842–854.
- [29] L. Xiao, P. Wang, D. Niyato, D. Kim, Z. Han, Wireless networks with RF energy harvesting: a contemporary survey, *IEEE Commun. Surveys Tuts.* 2 (17) (2015) 757–789.
- [30] C. Shen, W.-Q. Li, T.-H. Chang, Wireless information and power transfer with multi-antenna interference channel, *IEEE Trans. Sig. Proc.* 62 (23) (2014) 6249–6264.
- [31] S. Timotheou, I. Krikidis, G. Zheng, B. Ottersten, Wireless information and power transfer with multi-antenna interference channel, *IEEE Trans. Wireless Commun.* 13 (5) (2014) 2646–2658.
- [32] Q. Shi, W. Xu, T.H. Chang, Y. Wang, E. Song, Joint beamforming and power splitting for MISO interference channel with SWIPT: an SOCP relaxation and decentralized algorithm, *IEEE Trans. Sig. Proc.* 62 (23) (2014) 6194–6208.

- [33] S. Lee, L. Liu, R. Zhang, Collaborative wireless energy and information transfer in interference channel, *IEEE Trans. Wireless Commun.* 14 (1) (2015) 545–557.
- [34] L. Liu, R. Zhang, K.C. Chua, Multi-antenna wireless powered communication with energy beamforming, *IEEE Trans. Commun.* 62 (12) (2014) 4349–4361.
- [35] R. Horn, C. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.
- [36] G. Yang, C.K. Ho, R. Zhang, Y.L. Guan, Throughput optimization for massive MIMO systems powered by wireless energy transfer, *IEEE J. Sel. Areas Commun.* 33 (8) (2015) 1640–1650.
- [37] D.J. Love, R.W. Heath Jr, V.K.N. Lau, D. Gesbert, B.D. Rao, M. Andrews, An overview of limited feedback in wireless communication systems, *IEEE J. Sel. Areas Commun.* 26 (8) (2008) 1341–1365.
- [38] Y. Zeng, R. Zhang, Optimized training design for wireless energy transfer, *IEEE Trans. Commun.* 63 (2) (2015) 536–550.
- [39] G. Yang, C.K. Ho, Y.L. Guan, Dynamic resource allocation for multiple-antenna wireless power transfer, *IEEE Trans. Sig. Proc.* 62 (14) (2014) 3565–3577.
- [40] G. Yang, C.K. Ho, Y.L. Guan, Multi-antenna wireless energy transfer for backscatter communication system, *IEEE J. Sel. Areas Commun.* 12 (33) (2015) 2974–2987.

# Sparse methods for direction-of-arrival estimation

# 11

Zai Yang<sup>\*,†</sup>, Jian Li<sup>‡</sup>, Petre Stoica<sup>§</sup>, Lihua Xie<sup>†</sup>

Nanjing University of Science and Technology, Nanjing, China<sup>\*</sup> Nanyang Technological University, Singapore, Singapore<sup>†</sup>

University, Singapore<sup>†</sup> University of Florida, Gainesville, FL, USA<sup>‡</sup> Uppsala

University, Uppsala, Sweden<sup>§</sup>

## 11.1 INTRODUCTION

Direction-of-arrival (DOA) estimation refers to the process of retrieving the direction information of several electromagnetic waves/sources from the outputs of a number of receiving antennas that form a sensor array. DOA estimation is a major problem in array signal processing and has wide applications in radar, sonar, wireless communications, etc.

The study of DOA estimation methods has a long history. For example, the conventional (Bartlett) beamformer, which dates back to the World War II, simply uses Fourier-based spectral analysis of the spatially sampled data. Capon's beamformer was later proposed to improve the estimation performance of closely spaced sources [1]. Since the 1970s when Pisarenko found that the DOAs can be retrieved from data second order statistics [2], a prominent class of methods designated as subspace-based methods have been developed, e.g., the multiple signal classification (MUSIC) and the estimation of parameters by rotational invariant techniques (ESPRIT) along with their variants [3–7]. Another common approach is the nonlinear least squares (NLS) method that is also known as the (deterministic) maximum likelihood estimation. For a complete review of these methods, readers are referred to [8–10]. Note that these methods suffer from certain well-known limitations. For example, the subspace-based methods and the NLS need a priori knowledge on the source number that may be difficult to obtain. Additionally, Capon's beamformer, MUSIC and ESPRIT are covariance-based and require a sufficient number of data snapshots to accurately estimate the data covariance matrix. Moreover, they can be sensitive to source correlations that tend to cause a rank deficiency in the sample data covariance matrix. Also, a very accurate initialization is required for the NLS since its objective function has a complicated multimodal shape with a sharp global minimum.

The purpose of this article is to provide an overview of the recent work on sparse DOA estimation methods. These new methods are motivated by techniques in sparse representation and compressed sensing methodology [11–15], and most of them have been proposed during the last decade. The sparse estimation (or optimization) methods can be applied in several demanding scenarios, including cases with no knowledge of the source number, limited number of snapshots (even a single snapshot), and highly or completely correlated sources. Due to these attractive properties they have been extensively studied and their popularity is reflected by the large number of publications about them.

It is important to note that there is a key difference between the common sparse representation framework and DOA estimation. To be specific, the studies of sparse representation have been focused on *discrete linear* systems. In contrast to this, the DOA parameters are *continuous* valued and the observed data are *nonlinear* in the DOAs. Depending on the model adopted, we can classify the sparse methods for DOA estimation into three categories, namely, *on-grid*, *off-grid*, and *gridless*, which also corresponds to the chronological order in which they have been developed. For on-grid sparse methods, the data model is obtained by assuming that the true DOAs lie on a set of fixed grid points in order to straightforwardly apply the existing sparse representation techniques. While a grid is still required by off-grid sparse methods, the DOAs are not restricted to be on the grid. Finally, the recent gridless sparse methods do not need a grid, as their name suggests, and they operate directly in the continuous domain.

The organization of this article is as follows. The data model for DOA estimation is introduced in [Section 11.2](#) for far-field, narrowband sources that are the focus of this article. Its dependence on the array geometry and the parameter identifiability problem are discussed. In [Section 11.3](#) the concepts of sparse representation and compressed sensing are introduced and several sparse estimation techniques are discussed. Moreover, we discuss the feasibility of using sparse representation techniques for DOA estimation and highlight the key differences between sparse representation and DOA estimation. The on-grid sparse methods for DOA estimation are introduced in [Section 11.4](#). Since they are straightforward to obtain in the case of a single data snapshot, we focus on showing how the temporal redundancy of multiple snapshots can be utilized to improve the DOA estimation performance. Then, the off-grid sparse methods are presented in [Section 11.5](#). [Section 11.6](#) is the main highlight of this article in which the recently developed gridless sparse methods are presented. These methods are of great interest since they operate directly in the continuous domain and have strong theoretical guarantees. Some future research directions will be discussed in [Section 11.7](#) and conclusions will be drawn in [Section 11.8](#).

Notations used in this article are as follows.  $\mathbb{R}$  and  $\mathbb{C}$  denote the sets of real and complex numbers respectively. Boldface letters are reserved for vectors and matrices.  $|\cdot|$  denotes the amplitude of a scalar or the cardinality of a set.  $\|\cdot\|_1$ ,  $\|\cdot\|_2$ , and  $\|\cdot\|_F$  denote the  $\ell_1$ ,  $\ell_2$ , and Frobenius norms respectively.  $A^T$ ,  $A^*$ , and  $A^H$  are the matrix transpose, complex conjugate, and conjugate transpose of  $A$  respectively.  $x_j$  is the  $j$ th entry of a vector  $x$ , and  $A_j$  denotes the  $j$ th row of a matrix  $A$ . Unless

otherwise stated,  $\mathbf{x}_\Omega$  and  $\mathbf{A}_\Omega$  are the subvector and submatrix of  $\mathbf{x}$  and  $\mathbf{A}$  obtained by retaining the entries of  $\mathbf{x}$  and the rows of  $\mathbf{A}$  indexed by the set  $\Omega$ . For a vector  $\mathbf{x}$ ,  $\text{diag}(\mathbf{x})$  is a diagonal matrix with  $\mathbf{x}$  on the diagonal.  $\mathbf{x} \succeq \mathbf{0}$  means  $x_j \geq 0$  for all  $j$ .  $\text{rank}(\mathbf{A})$  denotes the rank of a matrix  $\mathbf{A}$  and  $\text{Tr}(\mathbf{A})$  denotes the trace. For positive semidefinite matrices  $\mathbf{A}$  and  $\mathbf{B}$ ,  $\mathbf{A} \geq \mathbf{B}$  means that  $\mathbf{A} - \mathbf{B}$  is positive semidefinite. Finally,  $\mathbb{E}[\cdot]$  denotes the expectation of a random variable, and for notational simplicity a random variable and its numerical value will not be distinguished.

---

## 11.2 DATA MODEL

### 11.2.1 DATA MODEL

In this section, the DOA estimation problem is stated. Consider  $K$  narrowband far-field source signals  $s_k$ ,  $k = 1, \dots, K$ , impinging on an array of omnidirectional sensors from directions  $\theta_k$ ,  $k = 1, \dots, K$ . According to [8, 9, 16], the time delays at different sensors can be represented by simple phase shifts, resulting in the following data model:

$$\mathbf{y}(t) = \sum_{k=1}^K \mathbf{a}(\theta_k) s_k(t) + \mathbf{e}(t) = \mathbf{A}(\boldsymbol{\theta}) \mathbf{s}(t) + \mathbf{e}(t), \quad t = 1, \dots, L, \quad (11.1)$$

where  $t$  indexes the snapshot and  $L$  is the number of snapshots,  $\mathbf{y}(t) \in \mathbb{C}^M$ ,  $\mathbf{s}(t) \in \mathbb{C}^K$ , and  $\mathbf{e}(t) \in \mathbb{C}^M$  denote the array output, the vector of source signals, and the vector of measurement noise at snapshot  $t$ , respectively, where  $M$  is the number of sensors.  $\mathbf{a}(\theta_k)$  is the so-called steering vector of the  $k$ th source that is determined by the geometry of the sensor array and will be given later. The steering vectors compose the array manifold matrix  $\mathbf{A}(\boldsymbol{\theta}) = [\mathbf{a}(\theta_1), \dots, \mathbf{a}(\theta_K)]$ . More compactly, Eq. (11.1) can be written as

$$\mathbf{Y} = \mathbf{A}(\boldsymbol{\theta}) \mathbf{S} + \mathbf{E}, \quad (11.2)$$

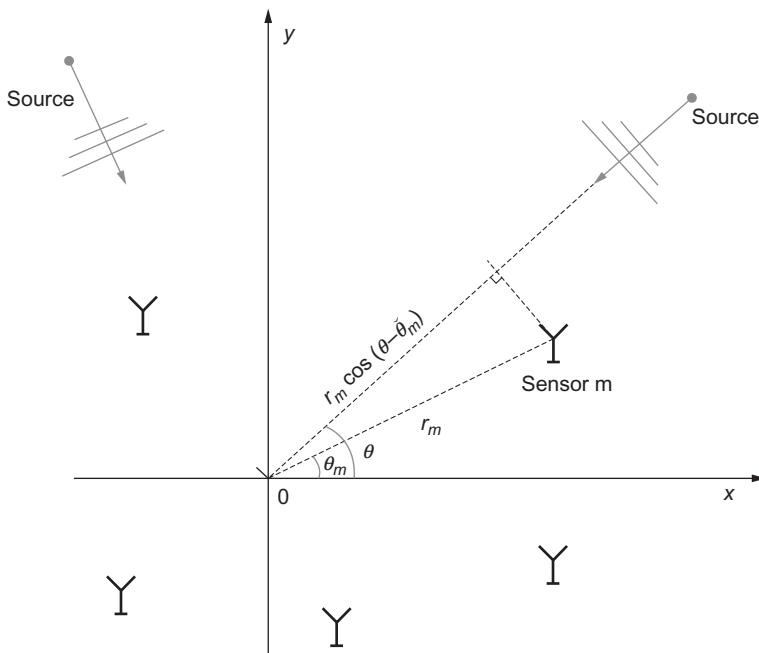
where  $\mathbf{Y} = [\mathbf{y}(1), \dots, \mathbf{y}(L)]$ , and  $\mathbf{S}$  and  $\mathbf{E}$  are similarly defined. Given the data matrix  $\mathbf{Y}$  and the mapping  $\boldsymbol{\theta} \rightarrow \mathbf{a}(\boldsymbol{\theta})$ , the objective is to estimate the parameters  $\theta_k$ ,  $k = 1, \dots, K$  that are referred to as the DOAs. It is worth noting that the source number  $K$  is usually unknown in practice; typically,  $K$  is assumed to be smaller than  $M$ , as otherwise the DOAs cannot be uniquely identified from the data (see details in [Section 11.2.3](#)).

### 11.2.2 THE ROLE OF ARRAY GEOMETRY

We now discuss how the mapping  $\boldsymbol{\theta} \rightarrow \mathbf{a}(\boldsymbol{\theta})$  is determined by the array geometry. We first consider a general 2-D array with the  $M$  sensors located at points  $(r_m, \theta_m)$ , expressed in polar coordinates (see [Fig. 11.1](#)). For convenience, the unit of distance is taken as half the wavelength of the waves. Then  $\mathbf{a}(\boldsymbol{\theta})$  will be given by

$$a_m(\theta) = e^{i\pi r_m \cos(\theta - \check{\theta}_m)}, \quad m = 1, \dots, M. \quad (11.3)$$

In the particularly interesting case of a linear array, assuming that the sensors are located on the nonnegative  $x$ -axis, we have that  $\check{\theta}_m = 0^\circ$ ,  $m = 1, \dots, M$ . Therefore,  $\mathbf{a}(\boldsymbol{\theta})$  will be given by

**FIG. 11.1**

The setup of the DOA estimation problem with a general 2-D array.

$$a_m(\theta) = e^{i\pi r_m \cos \theta}, \quad m = 1, \dots, M. \quad (11.4)$$

We can replace the variable  $\theta$  by  $f = \frac{1}{2} \cos \theta$  and define without ambiguity  $\mathbf{a}(f) = \mathbf{a}(\arccos(2f)) = \mathbf{a}(\theta)$ . Then, the mapping  $\mathbf{a}(f)$  is given by

$$a_m(f) = e^{i2\pi r_m f}, \quad m = 1, \dots, M. \quad (11.5)$$

In the case of a single snapshot, obviously, the *spatial* DOA estimation problem becomes a *temporal* frequency estimation problem (a.k.a. line spectral estimation) given the samples  $y_m, m = 1, \dots, M$  measured at time instants  $r_m, m = 1, \dots, M$ .

If we further assume that the sensors of the linear array are equally spaced, then the array is known as a uniform linear array (ULA) (see Fig. 11.2). We consider the case when two adjacent antennas of the array are spaced by a unit distance (half the wavelength). Then, we have that  $r_m = m - 1$  and

$$\mathbf{a}(f) = [1, e^{i2\pi f}, \dots, e^{i2\pi(M-1)f}]^T. \quad (11.6)$$

If a linear array is obtained from a ULA by retaining only a part of the sensors, then it is known as a sparse linear array (SLA).

It is worth noting that for a 2-D array, it is possible to estimate the DOAs in the entire  $360^\circ$  range, while for a linear array we can only resolve the DOAs in a  $180^\circ$

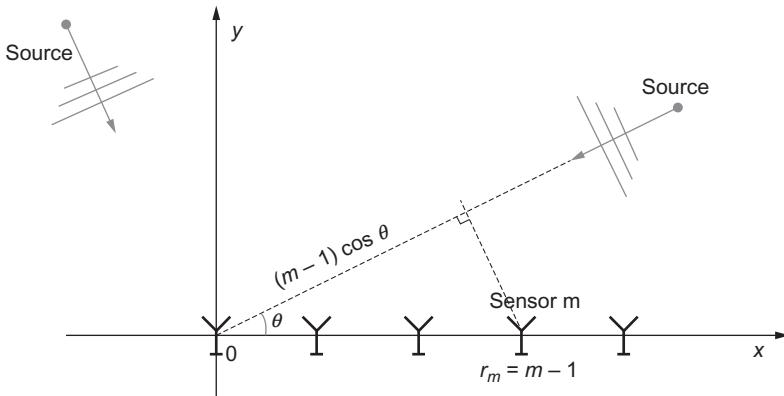


FIG. 11.2

The setup of the DOA estimation problem with a ULA.

range:  $\theta_k \in [0^\circ, 180^\circ]$ ). In the latter case, correspondingly, the “frequencies” are:  $f_k = \frac{1}{2} \cos \theta_k \in \left(-\frac{1}{2}, \frac{1}{2}\right]$ . Throughout this article, we let  $\mathcal{D}_\theta$  denote the domain of the DOAs that can be  $[0^\circ, 360^\circ]$  or  $[0^\circ, 180^\circ]$ , depending on the context. Also, let  $\mathbb{T} = \left(-\frac{1}{2}, \frac{1}{2}\right]$  be the frequency interval for linear arrays. Finally, we close this section by noting that the grid-based (on-grid and off-grid) sparse methods can be applied to arbitrary sensor arrays, while the existing gridless sparse methods are typically limited to ULAs or SLAs.

### 11.2.3 PARAMETER IDENTIFIABILITY

The DOAs  $\{\theta_k\}_{k=1}^K$  can be uniquely identified from  $\mathbf{Y} = \mathbf{A}(\boldsymbol{\theta})\mathbf{S}$  if there do not exist  $\{\theta'_k\}_{k=1}^K \neq \{\theta_k\}_{k=1}^K$  and  $\mathbf{S}'$  such that  $\mathbf{Y} = \mathbf{A}(\boldsymbol{\theta})\mathbf{S} = \mathbf{A}(\boldsymbol{\theta}')\mathbf{S}'$ . Guaranteeing that the parameters can be uniquely identified in the noiseless case is usually a prerequisite for their accurate estimation. The parameter identifiability problem for DOA estimation was studied in [17] for ULAs and in [18, 19] for general arrays. The results in [20–22] are also closely related to this problem. For a general array, define the set

$$\mathcal{A}_\theta = \{\mathbf{a}(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathcal{D}_\theta\}, \quad (11.7)$$

and define the spark of  $\mathcal{A}_\theta$ , denoted by  $\text{spark}(\mathcal{A}_\theta)$ , as the smallest number of elements in  $\mathcal{A}_\theta$  that are linearly dependent [23]. For any  $M$ -element array, it holds that

$$\text{spark}(\mathcal{A}_\theta) \leq M + 1. \quad (11.8)$$

Note that it is generally difficult to compute  $\text{spark}(\mathcal{A}_\theta)$ , except in the ULA case in which  $\text{spark}(\mathcal{A}_\theta) = M + 1$  by the fact that any  $M$  steering vectors in  $\mathcal{A}_\theta$  are linearly independent.

The paper [18] showed that any  $K$  sources can be uniquely identified from  $\mathbf{Y} = \mathbf{A}(\boldsymbol{\theta})\mathbf{S}$  provided that

$$K < \frac{\text{spark}(\mathcal{A}_\theta) - 1 + \text{rank}(\mathbf{S})}{2}. \quad (11.9)$$

Note that the above condition cannot be easily used in practice since it requires knowledge on  $\mathbf{S}$ . To resolve this problem, it was shown in [21] that the condition in Eq. (11.9) is equivalent to

$$K < \frac{\text{spark}(\mathcal{A}_\theta) - 1 + \text{rank}(\mathbf{Y})}{2}. \quad (11.10)$$

Moreover, the condition in Eq. (11.9) or (11.10) is also necessary [21]. Combining these results, we have the following theorem.

**Theorem 11.1** Any  $K$  sources can be uniquely identified from  $\mathbf{Y} = \mathbf{A}(\boldsymbol{\theta})\mathbf{S}$  if and only if the condition in Eq. (11.10) holds.

**Theorem 11.1** provides a necessary and sufficient condition for unique identifiability of the parameters. In the single snapshot case, the condition in Eq. (11.10) reduces to

$$K < \frac{\text{spark}(\mathcal{A}_\theta)}{2}. \quad (11.11)$$

Therefore, **Theorem 11.1** implies that more sources can be determined if more snapshots are collected, except in the trivial case when the source signals are identical up to scaling factors. In the ULA case, the condition in Eq. (11.10) can be simplified as

$$K < \frac{M + \text{rank}(\mathbf{Y})}{2}. \quad (11.12)$$

Using the inequality  $\text{rank}(\mathbf{S}) \leq K$  and Eq. (11.8), the condition in Eq. (11.9) or (11.10) implies that

$$K < \text{spark}(\mathcal{A}_\theta) - 1 \leq M. \quad (11.13)$$

**Theorem 11.1** specifies the condition required to guarantee unique identifiability for any  $K$  source signals. It was shown in [18] that if  $\{\theta_k\}$  are fixed and  $\mathbf{S}$  is randomly drawn from some absolutely continuous distribution, then the  $K$  sources can be uniquely identified with probability one, provided that

$$K < \frac{2\text{rank}(\mathbf{S})}{2\text{rank}(\mathbf{S}) + 1}(\text{spark}(\mathcal{A}_\theta) - 1). \quad (11.14)$$

Moreover, the following condition, which is slightly different from that in Eq. (11.14), is necessary:

$$K \leq \frac{2\text{rank}(\mathbf{S})}{2\text{rank}(\mathbf{S}) + 1}(\text{spark}(\mathcal{A}_\theta) - 1). \quad (11.15)$$

The condition in Eq. (11.14) is weaker than that in Eq. (11.9) or (11.10). As an example, in the single snapshot case, the upper bounds on  $K$  in Eqs. (11.10), (11.14) are

approximately  $\frac{1}{2}\text{spark}(\mathcal{A}_\theta)$  and  $\frac{2}{3}\text{spark}(\mathcal{A}_\theta)$ , respectively. However, the paper [19] pointed out that the condition in Eq. (11.14) has a relatively limited practical relevance in finite-SNR applications, since under Eq. (11.14), with a strictly positive probability, false DOA estimates far from the true DOAs may occur.

## 11.3 SPARSE REPRESENTATION AND DOA ESTIMATION

In this section we will introduce the basics of sparse representation that has been an active research topic especially in the past decade. More importantly, we will discuss its connections to and the key differences from DOA estimation.

### 11.3.1 SPARSE REPRESENTATION AND COMPRESSED SENSING

#### 11.3.1.1 Problem formulation

We first introduce the topic of sparse representation and the closely related concept of compressed sensing that have found broad applications in image, audio and signal processing, communications, medical imaging, and computational biology, to name just a few (see, e.g., the various special issues in several journals [24–27]). Let  $\mathbf{y} \in \mathbb{C}^M$  be the signal that we observe. We want to sparsely represent  $\mathbf{y}$  via the following model:

$$\mathbf{y} = \mathbf{Ax} + \mathbf{e}, \quad (11.16)$$

where  $\mathbf{A} \in \mathbb{C}^{M \times \bar{N}}$  is a *given* matrix, with  $M \ll \bar{N}$ , that is referred as a dictionary and whose columns are called atoms,  $\mathbf{x} \in \mathbb{C}^{\bar{N}}$  is a sparse coefficient vector (note that the notation  $N$  is reserved for later use), and  $\mathbf{e}$  accounts for the representation error. By sparsity we mean that only a few entries, say  $K \ll \bar{N}$ , of  $\mathbf{x}$  are nonzero and the rest are zero. This together with Eq. (11.16) implies that  $\mathbf{y}$  can be well approximated by a linear combination of  $K$  atoms in  $\mathbf{A}$ . The underlying motivation for the sparse representation is that even though the observed data  $\mathbf{y}$  lies in a high-dimensional space, it can actually be well approximated in some lower-dimensional subspace ( $K < M$ ). Given  $\mathbf{y}$  and  $\mathbf{A}$ , the problem of sparse representation is to find the sparse vector  $\mathbf{x}$  subject to data consistency.

The concept of sparse representation was later extended within the framework of compressed sensing [13–15]. In compressed sensing, a sparse signal, represented by the sparse vector  $\mathbf{x}$ , is recovered from undersampled linear measurements  $\mathbf{y}$ , i.e., the system model (11.16) applies with  $M \ll \bar{N}$ . In this context,  $\mathbf{y}$  is referred to as the compressive data,  $\mathbf{A}$  is the sensing matrix, and  $\mathbf{e}$  denotes the measurement noise. Note that a data model similar to Eq. (11.16) applies if the signal of interest is sparse in some domain. Given  $\mathbf{y}$  and  $\mathbf{A}$ , the problem of sparse signal recovery in compressed sensing is also to solve for the sparse vector  $\mathbf{x}$  subject to data consistency. With no rise for ambiguity we will not distinguish between the terminologies used for sparse representation and compressed sensing, as these two problems are very much alike.

To solve for the sparse signal, intuitively, we should find the sparsest solution. In the absence of noise, therefore, we should solve the following optimization problem:

$$\min_{\mathbf{x}} \|\mathbf{x}\|_0, \text{ subject to } \mathbf{y} = \mathbf{A}\mathbf{x}, \quad (11.17)$$

where  $\|\mathbf{x}\|_0 = \{j: x_j \neq 0\}$  counts the nonzero entries of  $\mathbf{x}$  and is referred to as the  $\ell_0$  (pseudo-)norm or the sparsity of  $\mathbf{x}$ . View  $\mathbf{A}$  as the set of its column vectors, and define its spark, denoted by  $\text{spark}(\mathbf{A})$  as in [Section 11.2.3](#). It can be shown that the true sparse signal  $\mathbf{x}$  can be uniquely determined by Eq. (11.17) if  $\mathbf{x}$  has a sparsity of

$$K < \frac{\text{spark}(\mathbf{A})}{2}. \quad (11.18)$$

To see this, suppose there exists  $\mathbf{x}'$  of sparsity  $K' \leq K$  satisfying  $\mathbf{y} = \mathbf{A}\mathbf{x}'$  as well. Then it holds that  $\mathbf{A}(\mathbf{x} - \mathbf{x}') = \mathbf{0}$ . Since  $\mathbf{x} - \mathbf{x}'$  has a sparsity of at most  $K + K' \leq 2K < \text{spark}(\mathbf{A})$ , which holds following Eq. (11.18), it can be concluded that  $\mathbf{x} - \mathbf{x}' = \mathbf{0}$  and thus  $\mathbf{x} = \mathbf{x}'$  since any  $\text{spark}(\mathbf{A}) - 1$  columns of  $\mathbf{A}$  are linearly independent. It is interesting to note that the condition in Eq. (11.18) is very similar to that in Eq. (11.11) required to guarantee identifiability for DOA estimation in the single snapshot case.

Unfortunately, the  $\ell_0$  optimization problem in Eq. (11.17) is NP hard to solve. Therefore, more efficient approaches are needed. We note that many methods and algorithms have been proposed for sparse signal recovery, e.g., convex relaxation or  $\ell_1$  optimization [11, 12],  $\ell_q$ ,  $0 < q < 1$  (pseudo-)norm optimization [28–34], greedy algorithms such as orthogonal matching pursuit (OMP), compressive sampling matching pursuit (CoSaMP) and subspace pursuit (SP) [35–40], iterative hard thresholding (IHT) [41], maximum likelihood estimation (MLE), etc. Readers can consult [42] for a review. Here we introduce convex relaxation,  $\ell_q$  optimization and MLE in the ensuing subsections.

### 11.3.1.2 Convex relaxation

The first practical approach to sparse signal recovery that we will introduce is based on the convex relaxation, which replaces the  $\ell_0$  norm by its tightest convex relaxation—the  $\ell_1$  norm. Therefore, we solve the following optimization problem in lieu of Eq. (11.17):

$$\min \|\mathbf{x}\|_1, \text{ subject to } \mathbf{y} = \mathbf{A}\mathbf{x}, \quad (11.19)$$

which is sometimes referred to as basis pursuit (BP) [11]. Since the  $\ell_1$  norm is convex, Eq. (11.19) can be solved in a polynomial time. In fact, the use of  $\ell_1$  optimization for obtaining a sparse solution dates back to the paper [43] about seismic data recovery. While the BP was empirically observed to give good performance, a rigorous analysis had not been provided for decades.

To introduce the existing theoretical guarantees for the BP in Eq. (11.19), we define a metric of the matrix  $\mathbf{A}$  called mutual coherence that quantifies the correlations between the atoms in  $\mathbf{A}$  [12].

**Definition 11.1** The mutual coherence of a matrix  $\mathbf{A}$ ,  $\mu(\mathbf{A})$ , is the largest absolute correlation between any two columns of  $\mathbf{A}$ , i.e.,

$$\mu(\mathbf{A}) = \max_{i \neq j} \frac{|\langle \mathbf{a}_i, \mathbf{a}_j \rangle|}{\|\mathbf{a}_i\|_2 \|\mathbf{a}_j\|_2}, \quad (11.20)$$

where  $\langle \cdot, \cdot \rangle$  denotes the inner product.

Intuitively, if two atoms in  $\mathbf{A}$  are highly correlated, then it will be difficult to distinguish their contributions to the measurements  $\mathbf{y}$ . In the extreme case when two atoms are completely coherent, it will be impossible to distinguish their contributions and thus impossible to recover the sparse signal  $\mathbf{x}$ . Therefore, to guarantee successful signal recovery, the mutual coherence  $\mu(\mathbf{A})$  should be small. This is true, according to the following theorem.

**Theorem 11.2** ([12]) Assume that  $\|\mathbf{x}\|_0 \leq K$  for the true signal  $\mathbf{x}$  and  $\mu < \frac{1}{2K-1}$ . Then,  $\mathbf{x}$  is the unique solution of the  $\ell_0$  optimization and the BP problem.

Another theoretical guarantee is based on the restricted isometry property (RIP) that quantifies the correlations of the atoms in  $\mathbf{A}$  in a different manner and has been popular in the development of compressed sensing.

**Definition 11.2** ([44]) The  $K$ -restricted isometry constant (RIC) of a matrix  $\mathbf{A}$ ,  $\delta_K(\mathbf{A})$ , is the smallest number such that the inequality

$$(1 - \delta_K(\mathbf{A}))\|\mathbf{v}\|_2^2 \leq \|\mathbf{A}\mathbf{v}\|_2^2 \leq (1 + \delta_K(\mathbf{A}))\|\mathbf{v}\|_2^2$$

holds for all  $K$ -sparse vectors  $\mathbf{v}$ .  $\mathbf{A}$  is said to satisfy the  $K$ -RIP with constant  $\delta_K(\mathbf{A})$  if  $\delta_K(\mathbf{A}) < 1$ .

By definition, matrices that have small RICs perform approximately orthogonal/unitary transformations when applied to sparse vectors. The following theoretical guarantee is provided in [45].

**Theorem 11.3** ([45]) Assume that  $\|\mathbf{x}\|_0 \leq K$  for the true signal  $\mathbf{x}$  and  $\delta_{2K} < \sqrt{2} - 1$ . Then  $\mathbf{x}$  is the unique solution of the  $\ell_0$  optimization and the BP problem.

After the work [45], the RIP condition has been improved, e.g., to  $\delta_{2K} < \frac{3}{4 + \sqrt{6}}$  [46]. Other types of RIP conditions are also available, e.g.,  $\delta_K < 0.307$  in [47]. It is known that stronger results can be provided by using RIP as compared to the mutual coherence. But it is worth noting that, unlike the mutual coherence that can be easily computed given the matrix  $\mathbf{A}$ , the complexity of computing the RIC of  $\mathbf{A}$  may increase dramatically with the sparsity  $K$ .

In the presence of noise we can solve the following regularized optimization problem, usually known as the least absolute shrinkage and selection operator (LASSO) [48]:

$$\min_{\mathbf{x}} \lambda \|\mathbf{x}\|_1 + \frac{1}{2} \|\mathbf{Ax} - \mathbf{y}\|_2^2, \quad (11.21)$$

where  $\lambda > 0$  is a regularization parameter, to be specified, or the basis pursuit denoising (BPDN) problem:

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1, \text{ subject to } \|\mathbf{Ax} - \mathbf{y}\|_2 \leq \eta, \quad (11.22)$$

where  $\eta \geq \|\mathbf{e}\|_2$  is an upper bound on the noise energy. Note that the problems in Eqs. (11.21), (11.22) are equivalent for appropriate choices of  $\lambda$  and  $\eta$ , and that both degenerate to BP in the noiseless case by letting  $\eta, \lambda \rightarrow 0$ . Under RIP conditions similar to the above ones, it has been shown that the sparse signal  $\mathbf{x}$  can be stably reconstructed with the reconstruction error being proportional to the noise level [45].

Besides the problems in Eqs. (11.21), (11.22), another  $\ell_1$  optimization method for sparse recovery is the so-called square-root LASSO [49]:

$$\min_{\mathbf{x}} \tau \|\mathbf{x}\|_1 + \|\mathbf{y} - \mathbf{Ax}\|_2, \quad (11.23)$$

where  $\tau > 0$  is a regularization parameter. Compared to the LASSO, for which the noise is usually assumed to be Gaussian and the regularization parameter  $\lambda$  is chosen proportional to the standard deviation of the noise, SR-LASSO requires a weaker assumption on the noise distribution and  $\tau$  can be chosen as a constant that is independent of the noise level [49].

The  $\ell_1$  optimization problems in Eqs. (11.19), (11.21), (11.22), (11.23) are convex and are guaranteed to be solvable in a polynomial time; however, it is not easy to efficiently solve them in the case when the problem dimension is high since the  $\ell_1$  norm is not a smooth function. Significant progress has been made over the past decade to accelerate the computation. Examples include  $\ell_1$ -magic [50], interior-point method [51], conjugate gradient method [52], fixed-point continuation [53], Nesterov's smoothing technique with continuation (NESTA) [54, 55], ONE-L1 algorithms [56], alternating direction method of multipliers (ADMM) [57, 58], and so on.

### 11.3.1.3 $\ell_q$ optimization

For a vector  $\mathbf{x}$ , the  $\ell_q$ ,  $0 < q < 1$  (pseudo-)norm is defined as:

$$\|\mathbf{x}\|_q = \left( \sum_n |x_n|^q \right)^{1/q}, \quad (11.24)$$

which is a nonconvex relaxation of the  $\ell_0$  norm. In lieu of Eqs. (11.17), (11.19), in the noiseless case, the  $\ell_q$  optimization problem is given by:

$$\min_{\mathbf{x}} \|\mathbf{x}\|_q^q, \text{ subject to } \mathbf{y} = \mathbf{Ax}, \quad (11.25)$$

where  $\|\mathbf{x}\|_q^q$  instead of  $\|\mathbf{x}\|_q$ , is used for the convenience of algorithm development. Since the  $\ell_q$  norm is a closer approximation to the  $\ell_0$  norm, compared to the  $\ell_1$  norm, it is expected that the  $\ell_q$  optimization in Eq. (11.25) results in better performance than the BP. This is true, according to [31, 33]. Indeed,  $\ell_q$ ,  $0 < q < 1$  optimization can exactly determine the true sparse signal under weaker RIP conditions than that for

the BP. Note that the results above are applicable to the globally optimal solution to Eq. (11.25), whereas we can only guarantee convergence to a locally optimal solution in practice.

A well-known algorithm for  $\ell_q$  optimization is the *focal underdetermined system solver* (FOCUSS) [28, 29]. FOCUSS is an iterative reweighted least squares method. In each iteration, FOCUSS solves the following weighted least squares problem:

$$\min_{\mathbf{x}} \sum_n w_n |x_n|^2, \text{ subject to } \mathbf{Ax} = \mathbf{y}, \quad (11.26)$$

where the weight coefficients  $w_n = |x_n|^{q-2}$  are updated using the latest solution  $\mathbf{x}$ . Note that Eq. (11.26) can be solved in closed form and hence an iterative algorithm can be implemented with a proper initialization. This algorithm can be interpreted as a majorization-minimization (MM) algorithm that is guaranteed to converge to a local minimum.

In the presence of noise, the following regularized problem is considered in lieu of Eq. (11.21):

$$\min_{\mathbf{x}} \lambda \|\mathbf{x}\|_q^q + \frac{1}{2} \|\mathbf{Ax} - \mathbf{y}\|_2^2, \quad (11.27)$$

where  $\lambda > 0$  is a regularization parameter. A regularized FOCUSS algorithm for Eq. (11.27) was developed in [30] by using the same main idea as in FOCUSS. A difficulty regarding Eq. (11.27) is the choice of the parameter  $\lambda$ . Although several heuristic methods for tuning this parameter were introduced in [30], to the best of our knowledge there have been no theoretical results on this aspect.

To bypass the parameter tuning problem, a maximum a posterior (MAP) estimation approach called SLIM (sparse learning via iterative minimization) was proposed in [34]. Assuming i.i.d. Gaussian noise with variance  $\eta$  and the following prior distribution for  $\mathbf{x}$ :

$$f(\mathbf{x}) \propto \prod_n e^{-\frac{2}{q}(|x_n|^q - 1)}, \quad (11.28)$$

SLIM computes the MAP estimate by solving the following  $\ell_q$  optimization problem:

$$\min_{\mathbf{x}} M \log \eta + \eta^{-1} \|\mathbf{Ax} - \mathbf{y}\|_2^2 + \frac{2}{q} \|\mathbf{x}\|_q^q. \quad (11.29)$$

To locally solve the problem in Eq. (11.29), SLIM iteratively updates  $\mathbf{x}$ , as the regularized FOCUSS does. However, unlike FOCUSS, SLIM also iteratively updates the parameter  $\eta$  based on the latest solution  $\mathbf{x}$ . Once  $q$  is given, SLIM is hyperparameter free. Note that Eq. (11.29) reduces to Eq. (11.27) for fixed  $\eta$ .

#### 11.3.1.4 Maximum likelihood estimation (MLE)

MLE is another common approach to sparse estimation. In contrast to the convex relaxation and OMP, one advantage of MLE is that it does not require knowledge of the noise level or the sparsity level (the latter being often needed to choose  $\lambda$  in Eq. 11.21 properly). To derive it, assume that  $\mathbf{x}$  follows a multivariate Gaussian

distribution with mean zero and covariance  $\mathbf{P} = \text{diag}(\mathbf{p})$ , where  $p_n \geq 0, p = 1, \dots, \bar{N}$  (this can be viewed as a prior distribution that does not necessarily have to hold in practice). Also, assume i.i.d. Gaussian noise with variance  $\sigma$ . It follows from the data model in Eq. (11.16) that  $\mathbf{y}$  follows a Gaussian distribution with mean zero and covariance  $\mathbf{R} = \mathbf{A}\mathbf{P}\mathbf{A}^H + \sigma\mathbf{I}$ . Consequently, the negative log-likelihood function associated with  $\mathbf{y}$  is given by

$$\mathcal{L}(\mathbf{p}, \sigma) = \log |\mathbf{R}| + \mathbf{y}^H \mathbf{R}^{-1} \mathbf{y}. \quad (11.30)$$

The parameters  $\mathbf{p}$  and  $\sigma$  can be estimated by minimizing  $\mathcal{L}$ :

$$\min_{\mathbf{p}, \sigma} \log |\mathbf{R}| + \mathbf{y}^H \mathbf{R}^{-1} \mathbf{y}. \quad (11.31)$$

Once  $\mathbf{p}$  and  $\sigma$  are solved for, the posterior distribution of the sparse signal  $\mathbf{x}$  can be obtained: it is a Gaussian distribution with mean and covariance given, respectively, by

$$\boldsymbol{\mu} = \boldsymbol{\Sigma} \mathbf{A}^H \mathbf{y}, \quad (11.32)$$

$$\boldsymbol{\Sigma} = (\mathbf{A}^H \mathbf{A} + \sigma \mathbf{P}^{-1})^{-1}. \quad (11.33)$$

The vector  $\mathbf{x}$  can be estimated as its posterior mean  $\boldsymbol{\mu}$ . In this process, the sparsity of  $\mathbf{x}$  is achieved by the fact that most of the entries of  $\mathbf{p}$  approach zero in practice. Theoretically, it can be shown that in the limiting noiseless case, the global optimizer to Eq. (11.31) coincides with that of  $\ell_0$  optimization [59].

The main difficulty of MLE is solving Eq. (11.31) in which the first term of the objective function, viz.  $\log |\mathbf{R}|$ , is a nonconvex (in fact, concave) function of  $(\mathbf{p}, \sigma)$ . Different approaches have been proposed, e.g., reweighted optimization [60] and sparse Bayesian learning (SBL) [59, 61, 62]. In [60], a majorization-minimization approach is adopted to linearize  $\log |\mathbf{R}|$  at each iteration by its tangent plane  $\text{Tr}(\mathbf{R}_j^{-1} \mathbf{R}) + \text{const}$  given the latest estimate  $\mathbf{R}_j$ . The resulting problem at each iteration is convex and solved using an algorithm called *sparse iterative covariance-based estimation* (SPICE) [60, 63–65] that will be introduced in Section 11.4.5.

The MLE has been interpreted from a different perspective within the framework of SBL or Bayesian compressed sensing. In particular, to achieve sparsity, a prior distribution is assumed for  $\mathbf{x}$  that promotes sparsity and is usually referred to as a sparse prior. In [61], for example, a Student's  $t$ -distribution is assumed for  $\mathbf{x}$  that is constructed in a hierarchical manner: specifically, a Gaussian distribution as above at the first stage followed by a Gamma distribution for the inverse of the powers,  $p_n^{-1}, n = 1, \dots, \bar{N}$  at the second stage. Interestingly, despite different approaches, the same objective function is obtained as in Eq. (11.31). To optimize Eq. (11.31), an expectation-maximization (EM) algorithm is adopted [66]. In the E-step, the posterior distribution of  $\mathbf{x}$  is computed, as mentioned previously, while in the M-step,  $\mathbf{p}$  and  $\sigma$  are updated as functions of the latest statistics of  $\mathbf{x}$ , viz.  $\boldsymbol{\mu}$  and  $\boldsymbol{\Sigma}$ . The process is repeated and it guarantees a monotonic decrease of  $\mathcal{L}$ . Finally, we note that with other sparse priors for  $\mathbf{x}$  that may possess different sparsity promoting properties, the obtained objective function of SBL can be slightly different from that of the MLE in Eq. (11.31) (see, e.g., [67]).

### 11.3.2 SPARSE REPRESENTATION AND DOA ESTIMATION: THE LINK AND THE GAP

In this subsection we discuss the link and the gap between sparse representation and DOA estimation. By doing so, we can see the possibility and the main challenges of using the sparse representation techniques for DOA estimation. It has been mentioned that the underlying motivation of sparse representation is that the observed data  $\mathbf{y}$  can be well approximated in a lower-dimensional subspace. In fact, this is exactly the case in DOA estimation where the data snapshot  $\mathbf{y}(t)$  is a linear combination of the steering vectors of the sources and the sparsity arises from the fact that there are less sources than sensors (note that for some special arrays and methods more sources than the sensors can be detected). By comparing the models in Eqs. (11.1), (11.16), it can be seen that the process of DOA estimation boils down to a sparse representation of the data snapshot with each DOA  $\theta$  corresponding to one atom given by  $\mathbf{a}(\theta)$ . Therefore, it is possible to use sparse representation techniques in DOA estimation.

However, there exist major differences between the common sparse representation framework and DOA estimation. First, and most importantly, the dictionary in sparse representation usually contains a finite number of atoms while in the DOA estimation problem the parameters are continuously valued, which leads to infinitely many atoms. More concretely, the atoms in sparse representation are given by the columns of a matrix. But in DOA estimation each atom  $\mathbf{a}(\theta)$  is parameterized by a continuous parameter  $\theta$ .

Second, there are usually multiple snapshots in DOA estimation problems, in contrast to the single snapshot case in sparse representation. It is then crucial to exploit the temporal redundancy of the snapshots in DOA estimation since the number of antennas can be limited due to physical and other constraints. Typically, the number of antennas  $M$  is about  $10 \sim 100$ , while the number of snapshots  $L$  can be much larger.

Last but not least, the existing theoretical guarantees of the sparse representation techniques are usually derived using what is known as incoherence analysis, e.g., those based on the mutual coherence and RIP, in the sense that they are applicable only in the case of incoherent dictionaries. This means that such guarantees can hardly be applied to DOA estimation problems, in which the atoms are completely coherent. But this does not necessarily mean that satisfactory performance cannot be achieved in DOA estimation problems. Indeed, note that the success of sparse signal recovery is measured by the size of the reconstruction error of the sparse signal  $\mathbf{x}$ , and that a slight error in the support usually results in a large estimation error. But this is not true for DOA estimation where the estimation error is actually measured by the error of the support (and, therefore, a small estimation error of the support is acceptable).

The next three sections describe three different possibilities for dealing with the first gap—discrete versus continuous atoms—when applying sparse representation to DOA estimation. In each section, we will also discuss how the signal sparsity

and the temporal redundancy of the multiple snapshots are exploited and what theoretical guarantees can be obtained.

## 11.4 ON-GRID SPARSE METHODS

In this section we introduce the first class of sparse methods for DOA estimation, termed as on-grid sparse methods. These methods are developed by directly applying sparse representation and compressed sensing techniques to DOA estimation. To do so, the DOAs are assumed to lie on a prescribed grid so that the problem can be solved within the common framework of sparse representation. The main challenge then is how to exploit the temporal redundancy of multiple snapshots.

In the following we first introduce the data model that we will use throughout this section. Then, we present several formulations and algorithms for DOA estimation within the on-grid framework, including  $\ell_{2,q}$  optimization methods with  $0 \leq q \leq 1$ , SBL and SPICE. Guidelines for grid selection will also be provided.

### 11.4.1 DATA MODEL

To fill the gap between continuous DOA estimation and discrete sparse representation, it is simply assumed by the on-grid sparse methods that the continuous DOA domain  $\mathcal{D}_\theta$  can be replaced by a *given* set of grid points

$$\bar{\boldsymbol{\theta}} = \{\bar{\theta}_1, \dots, \bar{\theta}_{\bar{N}}\}, \quad (11.34)$$

where  $\bar{N} \gg M$  is the grid size. This means that the candidate DOAs can only take values in  $\bar{\boldsymbol{\theta}}$ , which results in the following  $M \times \bar{N}$  dictionary matrix

$$\mathbf{A} = \mathbf{A}(\bar{\boldsymbol{\theta}}) = [\mathbf{a}(\bar{\theta}_1), \dots, \mathbf{a}(\bar{\theta}_{\bar{N}})]. \quad (11.35)$$

It follows that the data model in Eq. (11.2) for DOA estimation can be equivalently written as:

$$\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{E}, \quad (11.36)$$

where  $\mathbf{X} = [\mathbf{x}(1), \dots, \mathbf{x}(L)]$  is an  $\bar{N} \times L$  matrix in which each column  $\mathbf{x}(t)$  is an augmented version of the source signal  $s(t)$  and is defined by:

$$\mathbf{x}_n(t) = \begin{cases} s_k(t), & \text{if } \bar{\theta}_n = \theta_k; \\ 0, & \text{otherwise,} \end{cases} \quad n = 1, \dots, \bar{N}, t = 1, \dots, L. \quad (11.37)$$

It can be seen that for each  $t$ ,  $\mathbf{x}(t)$  contains only  $K$  nonzero entries, whose locations correspond to the  $K$  DOAs, and therefore it is a sparse vector as  $K \ll \bar{N}$ . Moreover,  $\mathbf{x}(t), t = 1, \dots, L$  are jointly sparse in the sense that they share the same support. Alternatively, we can say that  $\mathbf{X}$  is row-sparse in the sense that it contains only a few nonzero rows.

By means of the data model in Eq. (11.36), the DOA estimation problem is transformed into a sparse signal recovery problem. The DOAs are encoded in the support

of the sparse vectors  $\mathbf{x}(t)$ ,  $t = 1, \dots, L$  and therefore, we only need to recover this support from which the estimated DOAs can be retrieved.

The key and only difference between Eqs. (11.36) and (11.16) is that the former contains multiple data snapshots that are also referred to as multiple measurement vectors (MMVs). In the case of a single snapshot with  $L = 1$  (i.e., single measurement vector (SMV)), the sparse representation techniques can be readily applied to DOA estimation. In the case of multiple snapshots, the main difficulty consists in exploiting the temporal redundancy of the snapshots—the joint sparsity of the columns of  $\mathbf{X}$ —for possibly improved performance. Since the MMV data model in Eq. (11.36) is quite general, extensive studies have been performed for the joint sparse signal recovery problem (see, e.g., [20, 21, 64, 68–80]). We only discuss some of them in the ensuing subsections.

Before proceeding to the on-grid sparse methods, we make some comments on the data model in Eq. (11.36). Note that the set of grid points  $\bar{\theta}$  needs to be fixed a priori so that the dictionary  $\mathbf{A}$  is known, which is required in the sparse signal recovery process. Consequently, there is no guarantee that the true DOAs lie on the grid  $\bar{\theta}$ ; in fact, this fails with probability one, resulting in the grid mismatch problem [81, 82]. To ensure at least that the true DOAs are *close* to the grid points, in practice the grid needs to be dense enough (with  $\bar{N} \gg M$ ). Therefore, Eq. (11.36) can be viewed as a zeroth-order approximation of the true data model in Eq. (11.2) and the noise term  $\mathbf{E}$  in Eq. (11.36) may also comprise the approximation error besides the true noise in Eq. (11.2).

### 11.4.2 $\ell_{2,0}$ OPTIMIZATION

We first discuss how the joint sparsity can be exploited for sparse recovery. We start with the definition of sparsity for the row-sparse matrix  $\mathbf{X}$ . Since each row of  $\mathbf{X}$  corresponds to one potential source, it is natural to define the sparsity as the number of nonzero rows of  $\mathbf{X}$ , which is usually expressed as the following  $\ell_{2,0}$  norm (see, e.g., [20, 77]):

$$\|\mathbf{X}\|_{2,0} = \{n : \|\mathbf{X}_n\|_2 > 0\} = \{n : \mathbf{X}_n \neq \mathbf{0}\}, \quad (11.38)$$

where  $\mathbf{X}_n$  denotes the  $n$ th row of  $\mathbf{X}$ . Note that in Eq. (11.38) the  $\ell_2$  norm can in fact be replaced by any other norm. Following from the  $\ell_0$  optimization in the single snapshot case, the following  $\ell_{2,0}$  optimization can be proposed in the absence of noise:

$$\min_{\mathbf{X}} \|\mathbf{X}\|_{2,0}, \text{ subject to } \mathbf{Y} = \mathbf{AX}. \quad (11.39)$$

Suppose the optimal solution, denoted by  $\hat{\mathbf{X}}$ , can be obtained. Then, the DOAs can be retrieved from the row-support of  $\hat{\mathbf{X}}$ .

To realize the potential advantage of using the joint sparsity of the snapshots, consider the following result.

**Theorem 11.4** ([20]) *The true matrix  $\mathbf{X}$  is the unique solution to Eq. (11.39) if*

$$\|\mathbf{X}\|_{2,0} < \frac{\text{spark}(\mathbf{A}) - 1 + \text{rank}(\mathbf{Y})}{2}. \quad (11.40)$$

Note that the condition in Eq. (11.40) is very similar to that in Eq. (11.10) required to guarantee parameter identifiability for DOA estimation. By [Theorem 11.4](#), the number of recoverable DOAs can be increased in general by collecting more snapshots since then  $\text{rank}(\mathbf{Y})$  increases. The only exception happens in the case when the data snapshots  $\mathbf{y}(t)$ ,  $t = 1, \dots, L$  are identical up to scaling factors. Unfortunately, similar to the single snapshot case, the above  $\ell_{2,0}$  optimization problem is NP-hard to solve.

### 11.4.3 CONVEX RELAXATION

#### 11.4.3.1 $\ell_{2,1}$ optimization

The tightest convex relaxation of the  $\ell_{2,0}$  norm is given by the  $\ell_{2,1}$  norm that is defined as:

$$\|\mathbf{X}\|_{2,1} = \sum_n \|\mathbf{X}_n\|_2. \quad (11.41)$$

Though in the definition in Eq. (11.38) the  $\ell_2$  norm in the  $\ell_{2,0}$  norm can be replaced by other norms, its use is important in the  $\ell_{2,1}$  norm. Based on Eq. (11.39), the following  $\ell_{2,1}$  optimization problem is proposed in the absence of noise [68, 69]:

$$\min_{\mathbf{X}} \|\mathbf{X}\|_{2,1}, \text{ subject to } \mathbf{Y} = \mathbf{AX}. \quad (11.42)$$

As reported in the literature, the performance of  $\ell_{2,1}$  optimization approach can be generally improved by increasing the number of measurement vectors. Theoretically, this can be shown to be true under the assumption that the jointly sparse signals are randomly drawn such that the rows of the source signals  $\mathbf{S}_k$ ,  $k = 1, \dots, K$  are at general positions [76]. It is worth noting that the theoretical guarantee cannot be improved without assumptions on the source signals. To see this, consider the case when the columns of  $\mathbf{S}$  are identical up to scaling factors. Then, acquiring more snapshots does not provide useful information for DOA estimation. In this respect, the result of [76] can be referred to as *average case* analysis while those accounting for the aforementioned extreme case can be called *worst case* analysis.

In parallel to Eqs. (11.21), (11.22), in the presence of noise we can solve the LASSO problem:

$$\min_{\mathbf{X}} \lambda \|\mathbf{X}\|_{2,1} + \frac{1}{2} \|\mathbf{AX} - \mathbf{Y}\|_{\text{F}}^2, \quad (11.43)$$

where  $\lambda > 0$  is a regularization parameter (to be specified), or the BPDN problem:

$$\min_{\mathbf{X}} \|\mathbf{X}\|_{2,1}, \text{ subject to } \|\mathbf{AX} - \mathbf{Y}\|_{\text{F}} \leq \eta, \quad (11.44)$$

where  $\eta \geq \|\mathbf{E}\|_2$  is an upper bound on the noise energy. Note that it is generally difficult to choose  $\lambda$  in Eq. (11.43). Given the noise variance, results on choosing  $\lambda$  have recently been provided in [83–85] in the case of ULA and SLA. Readers are referred to [Section 11.6](#) for details.

Finally, we note that most, if not all, of the computational approaches to  $\ell_1$  optimization, e.g., those mentioned in [Section 11.3.1.2](#), can be easily extended to deal

with  $\ell_{2,1}$  optimization in the case of multiple snapshots. Once  $X$  is solved for, we can form a power spectrum by computing the power of each row of  $X$  from which the estimated DOAs can be obtained.

#### 11.4.3.2 Dimensionality reduction via $\ell_{2,1}$ -SVD

In DOA estimation applications the number of snapshots  $L$  can be large, which can significantly increase the computational workload of  $\ell_{2,1}$  optimization. In the case of  $L > K$  a dimensionality reduction technique was proposed in [69] inspired by the conventional subspace methods, e.g., MUSIC. In particular, suppose there is no noise; then the data snapshots  $Y$  lie in a  $K$ -dimensional subspace. In the presence of noise, therefore, we can decompose  $Y$  into the signal and noise subspaces, keep the signal subspace and use it in Eqs. (11.42)–(11.44) in lieu of  $Y$ . Mathematically, we compute the singular value decomposition (SVD)

$$Y = ULV^H. \quad (11.45)$$

Define a reduced  $M \times K$  dimensional data matrix

$$Y_{\text{SV}} = ULD_K^T = YVD_K^T \quad (11.46)$$

that contains most of the signal power, where  $D_K = [I_K, 0]$  with  $I_K$  being an identity matrix of order  $K$ . Also let  $X_{\text{SV}} = XVD_K^T$  and  $E_{\text{SV}} = EVD_K^T$ . Using these notations we can write a new data model:

$$Y_{\text{SV}} = AX_{\text{SV}} + E_{\text{SV}}. \quad (11.47)$$

Note that Eq. (11.47) is in exactly the same form as Eq. (11.36) but with reduced dimensionality. So similar  $\ell_{2,1}$  optimization problems can be formulated as Eqs. (11.43), (11.44), which are referred to as  $\ell_{2,1}$ -SVD.

The following comments on  $\ell_{2,1}$ -SVD are in order. Note that the true source number  $K$  has been known to obtain  $Y_{\text{SV}}$ . However,  $\ell_{2,1}$ -SVD is not very sensitive to this choice and therefore an appropriate estimate of  $K$  is sufficient in practice [69]. Nevertheless, parameter tuning remains a difficult problem for  $\ell_{2,1}$ -SVD ( $\lambda$  and  $\eta$  in Eqs. 11.43, 11.44). Though some solutions have been proposed for the standard  $\ell_{2,1}$  optimization methods in Eqs. (11.43), (11.44), given the noise level, they cannot be applied to  $\ell_{2,1}$ -SVD due to the change in the data structure. Regarding this aspect, it is somehow hard to compare the DOA estimation performances of the standard  $\ell_{2,1}$  optimization and  $\ell_{2,1}$ -SVD; though it is argued in [69] that, as compared to the standard form,  $\ell_{2,1}$ -SVD can improve the robustness to noise by keeping only the signal subspace.

#### 11.4.3.3 Another dimensionality reduction technique

We present here another dimensionality reduction technique that reduces the number of snapshots from  $L$  to  $M$  and has the same performance as the original  $\ell_{2,1}$  optimization. The technique was proposed in [86], inspired by a similar technique used for the gridless sparse methods (see Section 11.6). For convenience, we introduce it following the idea of  $\ell_{2,1}$ -SVD. In  $\ell_{2,1}$ -SVD the number of snapshots is reduced

from  $L$  to  $K$  by keeping only the  $K$ -dimensional signal subspace; in contrast the technique in [86] suggests keeping both the signal and noise subspaces. To be specific, suppose that  $L > M$  and  $\mathbf{Y}$  has rank  $r \leq M$  (note that typically  $r = M$  in the presence of noise). Then, given the SVD in Eq. (11.45), we retain a reduced  $M \times r$  dimensional data matrix

$$\mathbf{Y}_{\text{DR}} = \mathbf{U} \mathbf{L} \mathbf{D}_r^T = \mathbf{Y} \mathbf{V} \mathbf{D}_r^T \quad (11.48)$$

that preserves all of the data power since  $\mathbf{Y}$  has only  $r$  nonzero singular values, where  $\mathbf{D}_r$  is defined similarly to  $\mathbf{D}_K$ . We similarly define  $\mathbf{X}_{\text{DR}} = \mathbf{X} \mathbf{V} \mathbf{D}_r^T$  and  $\mathbf{E}_{\text{DR}} = \mathbf{E} \mathbf{V} \mathbf{D}_r^T$  to obtain the data model

$$\mathbf{Y}_{\text{DR}} = \mathbf{A} \mathbf{X}_{\text{DR}} + \mathbf{E}_{\text{DR}}. \quad (11.49)$$

For the LASSO problem, as an example, it can be shown that *equivalent* solutions can be obtained before and after the dimensionality reduction. To be specific, if  $\hat{\mathbf{X}}_{\text{DR}}$  is the solution to the following LASSO problem:

$$\min_{\mathbf{X}_{\text{DR}}} \lambda \|\mathbf{X}_{\text{DR}}\|_{2,1} + \frac{1}{2} \|\mathbf{A} \mathbf{X}_{\text{DR}} - \mathbf{Y}_{\text{DR}}\|_{\text{F}}^2, \quad (11.50)$$

then  $\hat{\mathbf{X}} = \hat{\mathbf{X}}_{\text{DR}} \mathbf{D}_r \mathbf{V}^H$  is the solution to Eq. (11.43).  $\hat{\mathbf{X}}_{\text{DR}}$  and  $\hat{\mathbf{X}}$  are equivalent in the sense that the corresponding rows of  $\hat{\mathbf{X}}_{\text{DR}}$  and  $\hat{\mathbf{X}}$  have the same power, resulting in identical power spectra (to see this, note that  $\hat{\mathbf{X}} \hat{\mathbf{X}}^H = \hat{\mathbf{X}}_{\text{DR}} \mathbf{D}_r \mathbf{V}^H \mathbf{V} \mathbf{D}_r^T \hat{\mathbf{X}}_{\text{DR}}^H = \hat{\mathbf{X}}_{\text{DR}} \hat{\mathbf{X}}_{\text{DR}}^H$ , whose diagonal contains the powers of the rows).

We next prove the above result. To do so, for any  $\mathbf{X}$ , split  $\mathbf{X} \mathbf{V}$  into two parts:  $\mathbf{X} \mathbf{V} = [\mathbf{X}_{\text{DR}}, \mathbf{X}_2]$ . Note that  $\mathbf{Y} \mathbf{V} = [\mathbf{Y}_{\text{DR}}, \mathbf{0}]$ . By the fact that  $\mathbf{V}$  is a unitary matrix, it can be easily shown that

$$\|\mathbf{X}\|_{2,1} = \|\mathbf{X} \mathbf{V}\|_{2,1} = \|[\mathbf{X}_{\text{DR}}, \mathbf{X}_2]\|_{2,1}, \quad (11.51)$$

$$\|\mathbf{A} \mathbf{X} - \mathbf{Y}\|_{\text{F}}^2 = \|\mathbf{A} \mathbf{X} \mathbf{V} - \mathbf{Y} \mathbf{V}\|_{\text{F}}^2 = \|\mathbf{A} \mathbf{X}_{\text{DR}} - \mathbf{Y}_{\text{DR}}\|_{\text{F}}^2 + \|\mathbf{A} \mathbf{X}_2\|_{\text{F}}^2. \quad (11.52)$$

It immediately follows that

$$\lambda \|\mathbf{X}\|_{2,1} + \frac{1}{2} \|\mathbf{A} \mathbf{X} - \mathbf{Y}\|_{\text{F}}^2 \geq \lambda \|\mathbf{X}_{\text{DR}}\|_{2,1} + \frac{1}{2} \|\mathbf{A} \mathbf{X}_{\text{DR}} - \mathbf{Y}_{\text{DR}}\|_{\text{F}}^2 \quad (11.53)$$

and the equality holds if and only if  $\mathbf{X}_2 = \mathbf{0}$ , or equivalently,  $\mathbf{X} = \mathbf{X}_{\text{DR}} \mathbf{D}_r \mathbf{V}^H$ . We can obtain the stated result by minimizing both sides of Eq. (11.53) with respect to  $\mathbf{X}$ .

Note that the above result also holds if  $\mathbf{Y}_{\text{DR}}$  is replaced by any full-column-rank matrix  $\tilde{\mathbf{Y}}$  satisfying  $\tilde{\mathbf{Y}} \tilde{\mathbf{Y}}^H = \mathbf{Y} \mathbf{Y}^H$ , since there always exists a unitary matrix  $\mathbf{V}$  such that  $\tilde{\mathbf{Y}} = \mathbf{Y} \mathbf{V} \mathbf{D}_r^T$  as for  $\mathbf{Y}_{\text{DR}}$ . Therefore, the SVD of the  $M \times L$  dimensional data matrix  $\mathbf{Y}$ , which can be computationally expensive in the case of  $L \gg M$ , can be replaced by the Cholesky decomposition or the eigenvalue decomposition of the  $M \times M$  matrix  $\mathbf{Y} \mathbf{Y}^H$  (which is the sample data covariance matrix up to a scaling factor). Another fact that makes this dimensional reduction technique superior to  $\ell_{2,1}$ -SVD is that the

parameter  $\lambda$  or  $\eta$  can be tuned as in the original  $\ell_{2,1}$  optimization, for which solutions are available if the noise level is given.

#### 11.4.4 $\ell_{2,q}$ OPTIMIZATION

Corresponding to the  $\ell_q$ ,  $0 < q < 1$  norm considered in Section 11.3.1.3, we can define the  $\ell_{2,q}$  norm to exploit the joint sparsity in  $\mathbf{X}$  as:

$$\|\mathbf{X}\|_{2,q} = \left( \sum_n \|\mathbf{X}_n\|_2^q \right)^{1/q}, \quad (11.54)$$

which is a nonconvex relaxation of the  $\ell_{2,0}$  norm. In lieu of Eqs. (11.42), (11.43), in the noiseless case, we can solve the following equality constrained problem:

$$\min_{\mathbf{X}} \|\mathbf{X}\|_{2,q}^q, \text{ subject to } \mathbf{A}\mathbf{X} = \mathbf{Y}, \quad (11.55)$$

or the following regularized form in the noisy case:

$$\min_{\mathbf{X}} \lambda \|\mathbf{X}\|_{2,q}^q + \frac{1}{2} \|\mathbf{A}\mathbf{X} - \mathbf{Y}\|_{\text{F}}^2. \quad (11.56)$$

To locally solve Eqs. (11.55), (11.56), the FOCUSS algorithm was extended in [68] to this multiple snapshot case to obtain M-FOCUSS. For Eq. (11.55), as in the single snapshot case, M-FOCUSS solves the following weighted least squares problem in each iteration:

$$\min_{\mathbf{X}} \sum_n w_n \|\mathbf{X}\|_2^2, \text{ subject to } \mathbf{A}\mathbf{X} = \mathbf{Y}, \quad (11.57)$$

where the weight coefficients  $w_n = \|\mathbf{X}\|_2^{q-2}$  are updated based on the latest solution  $\mathbf{X}$ . Since Eq. (11.57) can be solved in closed form, an iterative algorithm can be implemented. Note that Eq. (11.56) can be similarly solved as Eq. (11.55).

To circumvent the need for tuning the regularization parameter  $\lambda$  in Eq. (11.56), we can extend SLIM [34] to this multiple snapshot case as follows. Assume that the noise is i.i.d. Gaussian with variance  $\eta$  and that  $\mathbf{X}$  follows a prior distribution with the pdf given by

$$f(\mathbf{X}) \propto \prod_n e^{-\frac{2}{q}(\|\mathbf{X}_n\|_2^q - 1)}. \quad (11.58)$$

In Eq. (11.58), the  $\ell_2$  norm is performed on each row of  $\mathbf{X}$  to exploit the joint sparsity. As in the single snapshot case, SLIM computes the MAP estimator of  $\mathbf{X}$  which is the solution of the following  $\ell_{2,q}$  optimization problem:

$$\min_{\mathbf{x}} ML \log \eta + \eta^{-1} \|\mathbf{A}\mathbf{x} - \mathbf{Y}\|_{\text{F}}^2 + \frac{2}{q} \|\mathbf{x}\|_{2,q}^q. \quad (11.59)$$

Using a reweighting technique similar to that in M-FOCUSS, we can iteratively update  $\mathbf{X}$  and  $\eta$  in closed form and obtain the multiple snapshot version of SLIM. Finally, note

that the dimensionality reduction technique presented in Section 11.4.3.3 can also be applied to the  $\ell_{2,q}$  optimization problems in Eqs. (11.55), (11.56), (11.59) for algorithm acceleration [86].

### 11.4.5 SPARSE ITERATIVE COVARIANCE-BASED ESTIMATION (SPICE)

#### 11.4.5.1 Generalized least squares

To introduce SPICE, we first present the so-called generalized least squares method. To derive it, we need some statistical assumptions on the sources  $X$  and the noise  $E$ . We assume that  $\{\mathbf{x}(1), \dots, \mathbf{x}(L), \mathbf{e}(1), \dots, \mathbf{e}(L)\}$  are uncorrelated with one another and

$$\mathbb{E}\mathbf{e}(t)\mathbf{e}^H(t) = \sigma\mathbf{I}, \quad (11.60)$$

$$\mathbb{E}\mathbf{x}(t)\mathbf{x}^H(t) = \mathbf{P} = \text{diag}(\mathbf{p}), \quad t = 1, \dots, L, \quad (11.61)$$

where  $\sigma \geq 0$  and  $p_n \geq 0, n = 1, \dots, \bar{N}$  are the parameters of interest (note that the following derivations also apply to the case of heteroscedastic noise with  $\mathbb{E}\mathbf{e}(t)\mathbf{e}^H(t) = \text{diag}(\sigma_1, \dots, \sigma_M)$  with no or minor modifications). It follows that the snapshots  $\{\mathbf{y}(1), \dots, \mathbf{y}(L)\}$  are uncorrelated with one another and have the following covariance matrix:

$$\mathbf{R}(\mathbf{p}, \sigma) = \mathbb{E}\mathbf{y}(t)\mathbf{y}^H(t) = \mathbf{A}\mathbf{P}\mathbf{A}^H + \sigma\mathbf{I} = \mathbf{A}'\mathbf{P}'\mathbf{A}'^H, \quad (11.62)$$

where  $\mathbf{A}' = [\mathbf{A}, \mathbf{I}]$  and  $\mathbf{P}' = \text{diag}(\mathbf{P}, \sigma\mathbf{I})$ . Note that  $\mathbf{R}$  is linear in  $(\mathbf{p}, \sigma)$ . Let  $\tilde{\mathbf{R}} = \frac{1}{L}\mathbf{Y}\mathbf{Y}^H$  be the sample covariance matrix. Given  $\tilde{\mathbf{R}}$ , to estimate  $\mathbf{R}$  (in fact, the parameters  $\mathbf{p}$  and  $\sigma$  therein), we consider the generalized least squares method. First, we vectorize  $\tilde{\mathbf{R}}$  and let  $\tilde{\mathbf{r}} = \text{vec}(\tilde{\mathbf{R}})$  and  $\mathbf{r} = \text{vec}(\mathbf{R})$ . Since  $\tilde{\mathbf{R}}$  is an unbiased estimate of the data covariance matrix, it holds that

$$\mathbb{E}\tilde{\mathbf{r}} = \mathbf{r}. \quad (11.63)$$

Moreover, we can calculate the covariance matrix of  $\tilde{\mathbf{r}}$ , which is given by (see, e.g., [87])

$$\text{Cov}(\tilde{\mathbf{r}}) = \frac{1}{L}\mathbf{R}^T \otimes \mathbf{R}, \quad (11.64)$$

where  $\otimes$  denotes the Kronecker product. In the generalized least squares method we minimize the following criterion [87, 88]:

$$\begin{aligned} & \frac{1}{L}(\tilde{\mathbf{r}} - \mathbb{E}\tilde{\mathbf{r}})^H \text{Cov}^{-1}(\tilde{\mathbf{r}})(\tilde{\mathbf{r}} - \mathbb{E}\tilde{\mathbf{r}}) \\ &= (\tilde{\mathbf{r}} - \mathbf{r})^H [\mathbf{R}^{-T} \otimes \mathbf{R}^{-1}] (\tilde{\mathbf{r}} - \mathbf{r}) \\ &= \text{vec}^H(\tilde{\mathbf{R}} - \mathbf{R}) [\mathbf{R}^{-T} \otimes \mathbf{R}^{-1}] \text{vec}(\tilde{\mathbf{R}} - \mathbf{R}) \\ &= \text{vec}^H(\tilde{\mathbf{R}} - \mathbf{R}) \text{vec} \left\{ \mathbf{R}^{-1} (\tilde{\mathbf{R}} - \mathbf{R}) \mathbf{R}^{-1} \right\} \\ &= \text{Tr} \left\{ (\tilde{\mathbf{R}} - \mathbf{R}) \mathbf{R}^{-1} (\tilde{\mathbf{R}} - \mathbf{R}) \mathbf{R}^{-1} \right\} \\ &= \left\| \mathbf{R}^{-\frac{1}{2}} (\tilde{\mathbf{R}} - \mathbf{R}) \mathbf{R}^{-\frac{1}{2}} \right\|_F^2. \end{aligned} \quad (11.65)$$

The criterion in Eq. (11.65) has good statistical properties; for example, under certain conditions it provides a large-snapshot maximum likelihood (ML) estimator of the parameters  $(\boldsymbol{p}, \sigma)$  of interest. Unfortunately, Eq. (11.65) is nonconvex in  $\mathbf{R}$  and hence nonconvex in  $(\boldsymbol{p}, \sigma)$ . Therefore, there is no guarantee that it can be globally minimized.

Inspired by Eq. (11.65), the following convex criterion was proposed in [89]:

$$\left\| \tilde{\mathbf{R}}^{-\frac{1}{2}} (\tilde{\mathbf{R}} - \mathbf{R}) \tilde{\mathbf{R}}^{-\frac{1}{2}} \right\|_{\text{F}}^2, \quad (11.66)$$

in which  $\text{Cov}(\tilde{\mathbf{r}})$  in Eq. (11.64) is replaced by its consistent estimate, viz.  $\frac{1}{L} \tilde{\mathbf{R}}^T \otimes \tilde{\mathbf{R}}$ .

The resulting estimator remains a large-snapshot ML estimator. But it is only usable in the case of  $L \geq M$  when  $\tilde{\mathbf{R}}$  is nonsingular. The SPICE approach, which is discussed next, relies on Eq. (11.65) or (11.66) (see below for details).

### 11.4.5.2 SPICE

The SPICE algorithm has been proposed and studied in [60, 63–65]. In SPICE, the following covariance fitting criterion is adopted in the case of  $L \geq M$  whenever  $\tilde{\mathbf{R}}$  is nonsingular:

$$h_1 = \left\| \mathbf{R}^{-\frac{1}{2}} (\tilde{\mathbf{R}} - \mathbf{R}) \tilde{\mathbf{R}}^{-\frac{1}{2}} \right\|_{\text{F}}^2. \quad (11.67)$$

In the case of  $L < M$ , in which  $\tilde{\mathbf{R}}$  is singular, the following criterion is used instead:

$$h_2 = \left\| \mathbf{R}^{-\frac{1}{2}} (\tilde{\mathbf{R}} - \mathbf{R}) \right\|_{\text{F}}^2. \quad (11.68)$$

A simple calculation shows that

$$\begin{aligned} h_1 &= \text{Tr}(\mathbf{R}^{-1} \tilde{\mathbf{R}}) + \text{Tr}(\tilde{\mathbf{R}}^{-1} \mathbf{R}) - 2M \\ &= \text{Tr}\left(\tilde{\mathbf{R}}^{\frac{1}{2}} \mathbf{R}^{-1} \tilde{\mathbf{R}}^{\frac{1}{2}}\right) + \sum_{n=1}^N (\mathbf{a}_n^H \tilde{\mathbf{R}}^{-1} \mathbf{a}_n) p_n + \text{Tr}(\tilde{\mathbf{R}}^{-1}) \sigma - 2M. \end{aligned} \quad (11.69)$$

It follows that the optimization problem of SPICE based on  $h_1$  can be equivalently formulated as:

$$\min_{\mathbf{p} \succeq \boldsymbol{\theta}, \sigma > 0} \text{Tr}\left(\tilde{\mathbf{R}}^{\frac{1}{2}} \mathbf{R}^{-1} \tilde{\mathbf{R}}^{\frac{1}{2}}\right) + \sum_{n=1}^N (\mathbf{a}_n^H \tilde{\mathbf{R}}^{-1} \mathbf{a}_n) p_n + \text{Tr}(\tilde{\mathbf{R}}^{-1}) \sigma. \quad (11.70)$$

Note that the first term of the above objective function can be written as:

$$\text{Tr}\left(\tilde{\mathbf{R}}^{\frac{1}{2}} \mathbf{R}^{-1} \tilde{\mathbf{R}}^{\frac{1}{2}}\right) = \min \text{Tr}(X), \text{ subject to } \begin{bmatrix} \mathbf{X} & \tilde{\mathbf{R}}^{\frac{1}{2}} \\ \tilde{\mathbf{R}}^{\frac{1}{2}} & \mathbf{R} \end{bmatrix} \succeq \boldsymbol{\theta} \quad (11.71)$$

and hence it is convex in  $\mathbf{R}$  as well as in  $(\boldsymbol{p}, \sigma)$ . It follows that  $h_1$  is convex in  $(\boldsymbol{p}, \sigma)$ . Similarly, it holds for  $h_2$  that

$$\begin{aligned} h_2 &= \text{Tr}\left(\mathbf{R}^{-1}\tilde{\mathbf{R}}^2\right) + \text{Tr}(\mathbf{R}) - 2\text{Tr}\left(\tilde{\mathbf{R}}\right) \\ &= \text{Tr}\left(\tilde{\mathbf{R}}\mathbf{R}^{-1}\tilde{\mathbf{R}}\right) + \sum_{n=1}^{\bar{N}} \|\mathbf{a}_n\|_2^2 p_n + M\sigma - 2\text{Tr}\left(\tilde{\mathbf{R}}\right). \end{aligned} \quad (11.72)$$

The resulting optimization problem is given by:

$$\min_{\mathbf{p} \geq \mathbf{0}, \sigma > 0} \text{Tr}\left(\tilde{\mathbf{R}}\mathbf{R}^{-1}\tilde{\mathbf{R}}\right) + \sum_{n=1}^{\bar{N}} \|\mathbf{a}_n\|_2^2 p_n + M\sigma, \quad (11.73)$$

which is in a form similar to Eq. (11.70) and therefore is convex as well. Although both Eqs. (11.70), (11.73) can be cast as second order cone programs (SOCP) or semidefinite programs (SDP) (as shown above), for which standard solvers are available, they cannot be easily solved in practice based on these formulations due to the high dimensionality of the problem (note that  $\bar{N}$  can be very large).

We now introduce the SPICE algorithm to cope with the aforementioned computational problems. We focus on the case of  $L \geq M$  but similar results also hold in the case of  $L < M$ . The main result that underlies SPICE is the following reformulation (see, e.g., [64]):

$$\text{Tr}\left(\tilde{\mathbf{R}}^{\frac{1}{2}}\mathbf{R}^{-1}\tilde{\mathbf{R}}^{\frac{1}{2}}\right) = \min_{\mathbf{C}} \text{Tr}\left(\mathbf{C}^H \mathbf{P}'^{-1} \mathbf{C}\right), \text{ subject to } \mathbf{A}' \mathbf{C} = \tilde{\mathbf{R}}^{\frac{1}{2}} \quad (11.74)$$

and showing that the solution of  $\mathbf{C}$  is given by

$$\mathbf{C} = \mathbf{P}' \mathbf{A}'^H \mathbf{R}^{-1} \tilde{\mathbf{R}}^{\frac{1}{2}}. \quad (11.75)$$

Inserting Eq. (11.74) into Eq. (11.70), we see that the minimization of  $h_1$  can be equivalently written as:

$$\begin{aligned} \min_{\mathbf{C}, \mathbf{p} \geq \mathbf{0}, \sigma > 0} & \text{Tr}\left(\mathbf{C}^H \mathbf{P}'^{-1} \mathbf{C}\right) + \sum_{n=1}^{\bar{N}} \left( \mathbf{a}_n^H \tilde{\mathbf{R}}^{-1} \mathbf{a}_n \right) p_n + \text{Tr}\left(\tilde{\mathbf{R}}^{-1}\right) \sigma, \\ & \text{subject to } \mathbf{A}' \mathbf{C} = \tilde{\mathbf{R}}^{\frac{1}{2}}. \end{aligned} \quad (11.76)$$

Based on Eq. (11.76), the SPICE algorithm is derived by iteratively solving for  $\mathbf{C}$  and for  $(\mathbf{p}, \sigma)$ . First,  $\mathbf{p}$  and  $\sigma$  are initialized using, e.g., the conventional beamformer. Then,  $\mathbf{C}$  is updated using Eq. (11.75) with the latest estimates of  $\mathbf{p}$  and  $\sigma$ . After that, we update  $\mathbf{p}$  and  $\sigma$  by fixing  $\mathbf{C}$  and repeat the process until convergence. Note that  $(\mathbf{p}, \sigma)$  can also be determined in closed form, for fixed  $\mathbf{C}$ . To see this, observe that

$$\text{Tr}\left(\mathbf{C}^H \mathbf{P}'^{-1} \mathbf{C}\right) = \sum_{n=1}^{\bar{N}} \frac{\|\mathbf{C}_n\|_2^2}{p_n} + \frac{\sum_{n=\bar{N}+1}^{\bar{N}+M} \|\mathbf{C}_n\|_2^2}{\sigma}, \quad (11.77)$$

where  $\mathbf{C}_n$  denotes the  $n$ th row of  $\mathbf{C}$ . Inserting Eq. (11.77) in Eq. (11.76), the solutions  $p_n$ ,  $n = 1, \dots, \bar{N}$  and  $\sigma$  can be obtained as:

$$p_n = \frac{\|\mathbf{C}_n\|_2}{\sqrt{\mathbf{a}_n^H \tilde{\mathbf{R}}^{-1} \mathbf{a}_n}}, \quad n = 1, \dots, \bar{N}, \quad (11.78)$$

$$\sigma = \sqrt{\frac{\sum_{n=\bar{N}+1}^{\bar{N}+M} \|\mathbf{C}_n\|_2^2}{\text{Tr}(\tilde{\mathbf{R}}^{-1})}}. \quad (11.79)$$

Since the problem is convex and the objective function is monotonically decreasing in the iterative process, the SPICE algorithm is expected to converge to the global minimum. Note that the main computational cost of SPICE is for computing  $\mathbf{C}$  in Eqs. (11.78), (11.79) according to Eq. (11.75). It follows that SPICE has a per-iteration computational complexity of  $O(\bar{N}M^2)$  given that  $\bar{N} > M$ . Note also that, as compared to the original SPICE algorithm in [63, 64], a certain normalization step of the power estimates is removed here to avoid a global scaling ambiguity of the final power estimates (see also [60]).

We next discuss how the signal sparsity and joint sparsity are exploited in SPICE. Inserting Eqs. (11.78), (11.79) into Eq. (11.76), we see that the SPICE problem is equivalent to:

$$\min_{\mathbf{C}} \sum_{n=1}^{\bar{N}} \sqrt{\mathbf{a}_n^H \tilde{\mathbf{R}}^{-1} \mathbf{a}_n} \|\mathbf{C}_n\|_2 + \sqrt{\text{Tr}(\tilde{\mathbf{R}}^{-1}) \sum_{n=\bar{N}+1}^{\bar{N}+M} \|\mathbf{C}_n\|_2^2}, \quad (11.80)$$

subject to  $\mathbf{A}' \mathbf{C} = \tilde{\mathbf{R}}^{\frac{1}{2}}$ .

Note that the first term of the objective function in Eq. (11.80) is nothing but a weighted sum of the  $\ell_2$  norm of the first  $\bar{N}$  rows of  $\mathbf{C}$  (a.k.a. a weighted  $\ell_{2,1}$  norm) and thus promotes the row-sparsity of  $\mathbf{C}$ . Therefore, it is expected that most of  $\|\mathbf{C}_n\|_2, n = 1, \dots, \bar{N}$  will be equal to zero. This together with Eq. (11.78) implies that most of  $p_n, n = 1, \dots, \bar{N}$  will be zero and so sparsity is achieved. The joint sparsity is achieved by the assumption that the entries in each row of  $\mathbf{X}$  have identical variance  $p_n$ .

SPICE is related to the square-root LASSO in the single snapshot case. In particular, it was shown in [90, 91] that the SPICE problem in Eq. (11.73) is equivalent to

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1 + \|\mathbf{y} - \mathbf{Ax}\|_2, \quad (11.81)$$

which is nothing but the square-root LASSO in Eq. (11.23) with  $\tau = 1$ .

Finally, note that the decomposition in Eq. (11.62) is not unique in general (see Corollary 11.1 for detail). A direct consequence of this observation is that the SPICE algorithm generally does not provide unique estimates of  $\mathbf{p}$  and  $\sigma$  [92]. This problem will be fixed in the gridless version of SPICE that will be introduced in Sections 11.6.3.6 and 11.6.4.1.

### 11.4.6 MAXIMUM LIKELIHOOD ESTIMATION

The joint sparsity can also be exploited in the MLE, in a similar way as in SPICE. Assume that  $\mathbf{x}(t), t = 1, \dots, L$  are i.i.d. multivariate Gaussian distributed with mean zero and covariance  $\mathbf{P} = \text{diag}(\mathbf{p})$ . Also assume i.i.d. Gaussian noise with variance

$\sigma$  and that  $X$  and  $E$  are independent. Then, we have that the data snapshots  $y(t)$ ,  $t = 1, \dots, L$  are i.i.d. Gaussian distributed with mean zero and covariance  $\mathbf{R} = \mathbf{A}\mathbf{P}\mathbf{A}^H + \sigma\mathbf{I}$ . The negative log-likelihood function associated with  $\mathbf{Y}$  is therefore given by

$$\mathcal{L}(\mathbf{p}, \sigma) = \log |\mathbf{R}| + \text{Tr}(\mathbf{R}^{-1} \tilde{\mathbf{R}}) \quad (11.82)$$

where  $\tilde{\mathbf{R}}$  is the sample covariance matrix as defined in the preceding subsection. It follows that the parameters  $\mathbf{p}$  and  $\sigma$  can be estimated by solving the problem:

$$\min_{\mathbf{p}, \sigma} \log |\mathbf{R}| + \text{Tr}(\mathbf{R}^{-1} \tilde{\mathbf{R}}). \quad (11.83)$$

Owing to the analogy between Eq. (11.83) and its single snapshot counterpart (see Eq. 11.31), it should come as no surprise that the algorithms developed for the single snapshot case can also be applied to the multiple snapshot case with minor modifications. As an example, using a similar MM procedure, LIKES [60] can be extended to this multiple snapshot case.

The multiple snapshot MLE has been studied within the framework of SBL or Bayesian compressed sensing (see, e.g., [74, 93–95]). To exploit the joint sparsity of  $\mathbf{x}(t)$ ,  $t = 1, \dots, L$ , an identical sparse prior was assumed for all of them. The EM algorithm can also be used to perform parameter estimation via minimizing the objective in Eq. (11.83).

### 11.4.7 REMARKS ON GRID SELECTION

Based on the on-grid data model in Eq. (11.36), we have introduced several sparse optimization methods for DOA estimation in the preceding subsections. While we have focused on how the temporal redundancy or joint sparsity of the snapshots can be exploited, a major problem that remains unresolved is grid selection, i.e., the selection of the grid points  $\bar{\theta}$  in the data model (Eq. 11.36). Since discrete grid points are used to approximate the continuous DOA domain, intuitively, the grid should be chosen as fine as possible to improve the approximation accuracy. However, this can be problematic in two respects. Theoretically, a dense grid results in highly coherent atoms and hence few DOAs can be estimated according to the analysis based on the mutual coherence and RIP. Moreover, a too dense grid is not acceptable from an algorithmic viewpoint, since it will dramatically increase the computational complexity of an algorithm and also might cause slow convergence and numerical instability due to nearly identical adjacent atoms [96, 97].

To overcome the theoretical bottleneck mentioned above, the so-called coherence-inhibiting techniques have been proposed and incorporated in the existing sparse optimization methods to avoid solutions with closely located atoms [98, 99]. Additionally, with the development of recent gridless sparse methods it was shown that the local coherence between nearly located atoms actually does not matter for the convex relaxation approach if the true DOAs are appropriately separated [100] (details will be provided in Section 11.6).

To improve the computational speed and accuracy, a heuristic grid refinement strategy was proposed that suggests using a coarse grid at the initial stage and then gradually refining it based on the latest estimates of the DOAs [69]. A grid selection approach was also proposed by quantifying the similarity between the atoms in a grid bin [96]. In particular, suppose that in Eq. (11.36) the DOA interval  $(\theta_n - \frac{r}{2}, \theta_n + \frac{r}{2})$  is approximated by some grid point  $\theta_n$ , where  $r > 0$  denotes the grid interval. Then, on this interval the similarity is measured by the rank of the matrix defined by

$$\mathbf{C}_n = \int_{\theta_n - \frac{r}{2}}^{\theta_n + \frac{r}{2}} \mathbf{a}(v) \mathbf{a}^H(v) dv. \quad (11.84)$$

If  $\text{rank}(\mathbf{C}_n) \approx 1$ , then it is said that the grid is dense enough; otherwise, a denser grid is required. However, a problem with this criterion is that it can only be evaluated heuristically.

In summary, grid selection is an important problem that affects the practical DOA estimation accuracy, the computational speed and the theoretical analysis. A completely satisfactory solution to this problem within the framework of the on-grid methods seems hard to obtain since there always exist mismatches between the adopted discrete grid points and the true continuous DOAs.

## 11.5 OFF-GRID SPARSE METHODS

We have discussed the on-grid sparse methods in the preceding section, for which grid selection is a difficult problem and will inevitably result in grid mismatch. To resolve the grid mismatch problem, in this section we turn to the so-called off-grid sparse methods. In these methods, a grid is still required to perform sparse estimation but, unlike the on-grid methods, the DOA estimates are not restricted to be on the grid. We will mainly talk about two kinds of off-grid sparse methods: one is based on a fixed grid and joint estimation of the sparse signal and the grid offset, and the other relies on a dynamic grid. The main focus of the following discussions is on how to solve the grid mismatch.

### 11.5.1 FIXED GRID

#### 11.5.1.1 Data model

With a fixed grid  $\bar{\boldsymbol{\theta}} = \{\bar{\theta}_1, \dots, \bar{\theta}_N\}$ , an off-grid data model can be introduced as follows [101]. Suppose without loss of generality that  $\bar{\boldsymbol{\theta}}$  consists of uniformly spaced grid points with the grid interval  $r = \theta_2 - \theta_1 \propto \frac{1}{N}$ . For any DOA  $\theta_k$ , suppose  $\bar{\theta}_{n_k}$  is the nearest grid point with  $|\theta_k - \bar{\theta}_{n_k}| \leq \frac{r}{2}$ . We approximate the steering vector/atom  $\mathbf{a}(\theta_k)$  using a first-order Taylor expansion:

$$\mathbf{a}(\theta_k) \approx \mathbf{a}(\bar{\theta}_{n_k}) + \mathbf{b}(\bar{\theta}_{n_k})(\theta_k - \bar{\theta}_{n_k}), \quad (11.85)$$

where  $\mathbf{b}(\bar{\theta}_{n_k}) = \mathbf{a}'(\bar{\theta}_{n_k})$  (the derivative of  $\mathbf{a}(\theta)$ ). Similar to Eq. (11.36), we then obtain the following data model:

$$\mathbf{Y} = \Phi(\beta)\mathbf{X} + \mathbf{E}, \quad (11.86)$$

where  $\Phi(\beta) = \mathbf{A} + \mathbf{B} \operatorname{diag}(\beta)$ ,  $\mathbf{A} = [\mathbf{a}(\bar{\theta}_1), \dots, \mathbf{a}(\bar{\theta}_{\bar{N}})]$  is as defined previously,  $\mathbf{B} = [\mathbf{b}(\bar{\theta}_1), \dots, \mathbf{b}(\bar{\theta}_{\bar{N}})]$  and  $\beta = [\beta_1, \dots, \beta_{\bar{N}}] \in \left[-\frac{r}{2}, \frac{r}{2}\right]^{\bar{N}}$ , with

$$x_n(t) = \begin{cases} s_k(t), & \text{if } \bar{\theta}_n = \bar{\theta}_{n_k}; \\ 0, & \text{otherwise,} \end{cases} \quad (11.87)$$

$$\beta_n = \begin{cases} \theta_k - \bar{\theta}_{n_k}, & \text{if } \bar{\theta}_n = \bar{\theta}_{n_k}; \\ 0, & \text{otherwise,} \end{cases} \quad n = 1, \dots, \bar{N}, t = 1, \dots, L. \quad (11.88)$$

It follows from Eq. (11.86) that the DOA estimation problem can be formulated as sparse representation with uncertain parameters. In particular, once the row-sparse matrix  $\mathbf{X}$  and  $\beta$  can be estimated from  $\mathbf{Y}$ , then the DOAs can be estimated using the row-support of  $\mathbf{X}$  shifted by the offset  $\beta$ .

Compared to the on-grid model in Eq. (11.36), the additional grid offset parameters  $\beta_n$ ,  $n = 1, \dots, \bar{N}$  are introduced in the off-grid model in Eq. (11.86). Note that Eq. (11.86) reduces to Eq. (11.36) if  $\beta = \mathbf{0}$ . While Eq. (11.36) is based on a zeroth order approximation of the true data model, which causes grid mismatch, Eq. (11.86) can be viewed as a first-order approximation in which the grid mismatch can be partially compensated by jointly estimating the grid offset. Based on the off-grid model in Eq. (11.86), several methods have been proposed for DOA estimation by jointly estimating  $\mathbf{X}$  and  $\beta$  (see, e.g., [101–118]). Out of these methods we present the  $\ell_1$ -based optimization and SBL in the next subsections.

### 11.5.1.2 $\ell_1$ optimization

Inspired by the standard sparse signal recovery approach, several  $\ell_1$  optimization methods have been proposed to solve the off-grid DOA estimation problem. In [101], a sparse total least-squares (STLS) approach was proposed which, in the single snapshot case, solves the following LASSO-like problem:

$$\min_{\mathbf{x}, \beta} \lambda_1 \|\mathbf{x}\|_1 + \frac{1}{2} \|\mathbf{y} - [\mathbf{A} + \mathbf{B} \operatorname{diag}(\beta)]\mathbf{x}\|_2^2 + \lambda_2 \|\beta\|_2^2, \quad (11.89)$$

where  $\lambda_1$  and  $\lambda_2$  are regularization parameters. In Eq. (11.89), the prior information that  $\beta \in \left[-\frac{r}{2}, \frac{r}{2}\right]^{\bar{N}}$  is not used. To heuristically control the magnitude of  $\beta$ , its power is also minimized. Note that the problem in Eq. (11.89) is nonconvex due to the bilinear term  $\operatorname{diag}(\beta)\mathbf{x}$ . To solve Eq. (11.89), an alternating algorithm is adopted, iteratively solving for  $\mathbf{x}$  and  $\beta$ . Moreover, Eq. (11.89) can be easily extended to the multiple snapshot case by using  $\ell_{2,1}$  optimization to exploit the joint sparsity in  $\mathbf{X}$  (as in the preceding section). A difficult problem of these methods is parameter tuning, i.e., how to choose  $\lambda_1$  and  $\lambda_2$ .

To exploit the prior knowledge that  $\beta \in \left[-\frac{r}{2}, \frac{r}{2}\right]^N$ , the following BPDN-like formulation was proposed in the single snapshot case [103]:

$$\min_{x, \beta \in \left[-\frac{r}{2}, \frac{r}{2}\right]^N} \|x\|_1, \text{ subject to } \|y - [\mathbf{A} + \mathbf{B} \operatorname{diag}(\beta)]x\|_2 \leq \eta. \quad (11.90)$$

Note that Eq. (11.90) can be easily extended to the multiple snapshot case by using the  $\ell_{2,1}$  norm. In Eq. (11.90),  $\eta$  can be set according to information about the noise level and a possible estimate of the modeling error. Similar to Eq. (11.89), Eq. (11.90) is nonconvex, and a similar alternating algorithm can be implemented to monotonically decrease the value of the objective function. Note that if  $\beta$  is initialized as a zero vector, then the first iteration coincides with the standard BPDN.

It was shown in [103] that if the matrix  $[\mathbf{A}, \mathbf{B}]$  satisfies a certain RIP condition, then both  $x$  and  $\beta$  can be stably reconstructed, as in the standard sparse representation problem, with the reconstruction error being proportional to the noise level  $\eta$ . This means that in the ideal case of  $\eta = 0$  (assuming there is no noise or modeling error),  $x$  and  $\beta$  can be exactly recovered. A key step in showing this result is reformulating Eq. (11.90) as

$$\min_{x, \beta \in \left[-\frac{r}{2}, \frac{r}{2}\right]^N} \|x\|_1, \text{ subject to } \left\| y - [\mathbf{A} \ \mathbf{B}] \begin{bmatrix} x \\ \beta \odot x \end{bmatrix} \right\|_2 \leq \eta, \quad (11.91)$$

where  $\odot$  denotes the element-wise product. Although the RIP condition cannot be easily applied to the case of dense grid, the aforementioned result implies, to some extent, the superior performance of this off-grid optimization method as compared to the on-grid approach.

Following the lead of [103], a convex optimization method was proposed in [105] by exploiting the joint sparsity of  $x$  and  $v = \beta \odot x$ . In particular, the following problem was formulated:

$$\min_{x, v} \lambda \| [x \ v] \|_{2,1} + \frac{1}{2} \left\| y - [\mathbf{A} \ \mathbf{B}] \begin{bmatrix} x \\ v \end{bmatrix} \right\|_2^2, \quad (11.92)$$

which is equivalent to the following problem, for appropriate parameter choices:

$$\min_{x, v} \| [x \ v] \|_{2,1}, \text{ subject to } \left\| y - [\mathbf{A} \ \mathbf{B}] \begin{bmatrix} x \\ v \end{bmatrix} \right\|_2 \leq \eta. \quad (11.93)$$

This approach is advantageous in that it is convex and can be globally solved in a polynomial time, with similar theoretical guarantees as provided in [103]. However, it is worth noting that the prior knowledge on  $\beta$  cannot be exploited in this method. Additionally, the obtained solution for  $\beta_n = \frac{v_n}{x_n}$  might not even be real. To resolve this problem, Tan et al. [105] suggests a two-stage solution: (1) obtain  $x$  from Eq. (11.92), and (2) fix  $x$  and solve for  $\beta$  by minimizing  $\|y - [\mathbf{A} + \mathbf{B} \operatorname{diag}(\beta)]x\|_2$ .

### 11.5.1.3 Sparse Bayesian learning

A systematic approach to off-grid DOA estimation, called off-grid sparse Bayesian inference (OGSBI), was proposed in [104] within the framework of SBL in the multiple snapshot case. In order to estimate the additional parameter  $\beta$ , it is assumed that  $\beta_n, n = 1, \dots, \bar{N}$  are i.i.d. uniformly distributed on the interval  $[-\frac{r}{2}, \frac{r}{2}]$ . In the resulting EM algorithm, the posterior distribution of the row-sparse signal  $X$  can be computed in the expectation step as in the standard SBL. In the maximization step,  $\beta$  is also updated, in addition to updating the power  $p$  of the row-sparse signal and the noise variance  $\sigma$ . As in the standard SBL, the likelihood is guaranteed to monotonically increase and hence convergence of the algorithm can be obtained.

## 11.5.2 DYNAMIC GRID

### 11.5.2.1 Data model

The data model now uses a dynamic grid  $\bar{\theta}$  in the sense that the grid points  $\theta_n, n = 1, \dots, \bar{N}$  are not fixed:

$$\mathbf{Y} = \mathbf{A}(\bar{\theta})\mathbf{X} + \mathbf{E}. \quad (11.94)$$

For this model we need to jointly estimate the row-sparse matrix  $X$  and the grid  $\bar{\theta}$ . Once they are obtained, the DOAs are estimated using those grid points of  $\bar{\theta}$  corresponding to the nonzero rows of  $X$ . Since  $\bar{\theta}_n$ 's are estimated from the data and can be any values in the continuous DOA domain, this off-grid data model is accurate and does not suffer from grid mismatch. However, the difficulty lies in designing an algorithm for the joint estimation of  $X$  and  $\bar{\theta}$ , due to the nonlinearity of the mapping  $\mathbf{a}(\theta)$ . Note that the following algorithms that we will introduce are designated as off-grid methods, instead of gridless, since grid selection remains involved in them (e.g., choice of  $\bar{N}$  and initialization of  $\bar{\theta}$ ), which affects the computational speed and accuracy of the algorithms.

### 11.5.2.2 Algorithms

Several algorithms have been proposed based on the data model in Eq. (11.94). The first class is within the framework of SBL (see, e.g., [119–122]). But instead of using the EM algorithm as previously, a variational EM algorithm (or variational Bayesian inference) is typically exploited to carry out the sparse signal and parameter estimation. The reason is that the posterior distribution of the sparse vector  $x$  usually cannot be explicitly computed here, and that distribution is required by the EM but not by the variational EM. The main difficulty of these algorithms is the update of  $\theta$  due to the strong nonlinearity. Because closed-form solutions are not available, only numerical approaches can be used.

Another class of methods uses  $\ell_1$  optimization. In the single snapshot case, as an example, the paper [97] used a small  $\bar{N} \geq K$  and attempted to solve the following  $\ell_1$  optimization problem by iteratively updating  $x$  and  $\theta$ :

$$\min_{x, \theta} \lambda \|x\|_1 + \frac{1}{2} \|\mathbf{y} - \mathbf{A}(\bar{\theta})x\|_2^2. \quad (11.95)$$

To avoid the possible convergence of some  $\bar{\theta}_n$ 's to the same value, an additional (nonconvex) term  $g(\bar{\theta})$  is included to penalize closely located parameters:

$$\min_{\mathbf{x}, \bar{\theta}} \lambda_1 \|\mathbf{x}\|_1 + \frac{1}{2} \|\mathbf{y} - \mathbf{A}(\bar{\theta})\mathbf{x}\|_2^2 + \lambda_2 g(\bar{\theta}), \quad (11.96)$$

where  $\lambda_1$  and  $\lambda_2$  are regularization parameters that need to be tuned. Note that both Eqs. (11.95) and (11.96) are nonconvex. Even for given  $\mathbf{x}$ , it is difficult to solve for  $\bar{\theta}$ . Moreover, parameter tuning is tricky. Note that  $\ell_q$ ,  $q < 1$  optimization was also considered in [97] to enhance sparsity but it suffers from similar problems.

To promote sparsity, similar to  $\ell_1$  optimization, the following problem was proposed in [123, 124]:

$$\min_{\mathbf{x}, \bar{\theta}} \sum_{n=1}^N \lambda \log(|x_n|^2 + \epsilon) + \|\mathbf{y} - \mathbf{A}(\bar{\theta})\mathbf{x}\|_2^2. \quad (11.97)$$

To locally solve Eq. (11.97),  $\mathbf{x}$  and  $\bar{\theta}$  are iteratively updated. To solve for  $\mathbf{x}$  in closed form, the first term of the objective in Eq. (11.97) is replaced by a quadratic surrogate function that guarantees the decrease of the objective. The gradient descent method is then used to solve for  $\bar{\theta}$ . While it is generally difficult to choose  $\lambda$ , Fang et al. [124] suggested setting  $\lambda$  proportional to the inverse of the noise variance, leaving a constant coefficient to be tuned.

To conclude, in this section we introduced several off-grid sparse optimization methods for DOA estimation. By adopting a dynamic grid or estimating the grid offset jointly with the sparse signal, the grid mismatch encountered in the on-grid sparse methods can be overcome. However, this introduces more variables that need to be estimated and complicates the algorithm design. As a consequence, most of the presented algorithms involve nonconvex optimization and thus only local convergence can be guaranteed (except for the algorithm in [105]). Moreover, few theoretical guarantees can be obtained for most of the algorithms (see, however, [103, 105]).

## 11.6 GRIDLESS SPARSE METHODS

In this section, we present several recent DOA estimation approaches that are designated as the gridless sparse methods. As their name suggests, these methods do not require gridding of the direction domain. Instead, they directly operate in the continuous domain and therefore can completely resolve the grid mismatch problem. Moreover, they are convex and have strong theoretical guarantees. However, so far, this kind of methods can only be applied to uniform or sparse linear arrays. Therefore, naturally, in this section we treat the DOA estimation problem as frequency estimation, following the discussions in Section 11.2.

The rest of this section is organized as follows. We first revisit the data model in the context of ULAs and SLAs. We then introduce a mathematical tool known as the Vandermonde decomposition of Toeplitz covariance matrices, which is crucial for most gridless sparse methods. Finally, we discuss a number of gridless sparse

methods for DOA/frequency estimation in the case of a single snapshot, followed by the case of multiple snapshots. The atomic norm and gridless SPICE methods will be particularly highlighted.

### 11.6.1 DATA MODEL

For convenience, we restate the data model that will be used in this section. For an  $M$ -element ULA, the array data are modeled as:

$$\mathbf{Y} = \mathbf{A}(\mathbf{f})\mathbf{S} + \mathbf{E}, \quad (11.98)$$

where  $f_k \in \mathbb{T}$ ,  $k = 1, \dots, K$  are the frequencies of interest, which have a one-to-one relationship to the DOAs,  $\mathbf{A}(\mathbf{f}) = [\mathbf{a}(f_1), \dots, \mathbf{a}(f_K)] \in \mathbb{C}^{M \times K}$  is the array manifold matrix, and  $\mathbf{a}(\mathbf{f}) = [1, e^{j2\pi f}, \dots, e^{j2\pi(M-1)f}]^T \in \mathbb{C}^M$  is the steering vector.

For an  $M$ -element SLA, suppose that the array is obtained from an  $N$ -element virtual ULA by retaining the antennas indexed by the set  $\Omega = \{\Omega_1, \dots, \Omega_M\}$ , where  $N \geq M$  and  $1 \leq \Omega_1 < \dots < \Omega_M \leq N$ . In this case, we can view Eq. (11.98) as the data model with the virtual ULA. Then the data model of the SLA is given by

$$\mathbf{Y}_\Omega = \mathbf{A}_\Omega(\mathbf{f})\mathbf{S} + \mathbf{E}_\Omega. \quad (11.99)$$

Therefore, Eq. (11.98) can be considered as a special case of Eq. (11.99) in which  $M = N$  and  $\Omega = \{1, \dots, N\}$ . Given  $\mathbf{Y}$  (or  $\mathbf{Y}_\Omega$  and  $\Omega$ ), the objective is to estimate the frequencies  $f_k$ 's (note that the source number  $K$  is usually unknown).

In the single snapshot case the above problem coincides with the line spectral estimation problem. Since the first gridless sparse methods were developed for the latter problem, we present them in the single snapshot case and then discuss how they can be extended to the multiple snapshot case by exploiting the joint sparsity of the snapshots. Before doing that, an important mathematical tool is introduced in the following subsection.

### 11.6.2 VANDERMONDE DECOMPOSITION OF TOEPLITZ COVARIANCE MATRICES

The Vandermonde decomposition of Toeplitz covariance matrices plays an important role in this section. This classical result was discovered by Carathéodory and Fejér in 1911 [125]. It has become important in the area of data analysis and signal processing since the 1970s when it was rediscovered by Pisarenko and used for frequency retrieval from the data covariance matrix [2]. From then on, the Vandermonde decomposition has formed the basis of a prominent subset of methods for frequency and DOA estimation, viz. the subspace-based methods. To see why it is so, let us consider the data model in Eq. (11.98) and assume uncorrelated sources. In the noiseless case, the data covariance matrix is given by

$$\mathbf{R} = \mathbb{E}\mathbf{y}(t)\mathbf{y}^H(t) = \mathbf{A}(\mathbf{f})\text{diag}(\mathbf{p})\mathbf{A}^H(\mathbf{f}), \quad (11.100)$$

where  $p_k > 0$ ,  $k = 1, \dots, K$  are the powers of the sources. It can be easily verified that  $\mathbf{R}$  is a (Hermitian) Toeplitz matrix that can be written as:

$$\mathbf{R} = \mathbf{T}(\mathbf{u}) = \begin{bmatrix} u_1 & u_2 & \cdots & u_N \\ u_2^* & u_1 & \cdots & u_{N-1} \\ \vdots & \vdots & \ddots & \vdots \\ u_N^* & u_{N-1}^* & \cdots & u_1 \end{bmatrix}, \quad (11.101)$$

where  $\mathbf{u} \in \mathbb{C}^N$ . Moreover,  $\mathbf{R}$  is PSD and has rank  $K$  under the assumption that  $K < N$ . The Vandermonde decomposition result states that any rank-deficient, PSD Toeplitz matrix  $\mathbf{T}$  can be uniquely decomposed as in Eq. (11.100). Equivalently stated, this means that the frequencies can be exactly retrieved from the data covariance matrix. Formally, the result is stated in the following theorem, a proof of which (inspired by Gurvits and Barnum [126]) is also provided; note that the proof suggests a way of computing the decomposition.

**Theorem 11.5** Any PSD Toeplitz matrix  $\mathbf{T}(\mathbf{u}) \in \mathbb{C}^{N \times N}$  of rank  $r \leq N$  admits the following  $r$ -atomic Vandermonde decomposition:

$$\mathbf{T} = \sum_{k=1}^r p_k \mathbf{a}(f_k) \mathbf{a}^H(f_k) = \mathbf{A}(\mathbf{f}) \text{diag}(\mathbf{p}) \mathbf{A}^H(\mathbf{f}), \quad (11.102)$$

where  $p_k > 0$ , and  $f_k \in \mathbb{T}$ ,  $k = 1, \dots, r$  are distinct. Moreover, the decomposition in Eq. (11.102) is unique if  $r < N$ .

*Proof.* We first consider the case of  $r = \text{rank}(\mathbf{T}) \leq N - 1$ . Since  $\mathbf{T} \geq \mathbf{0}$ , there exists  $\mathbf{V} \in \mathbb{C}^{N \times r}$  satisfying  $\mathbf{T} = \mathbf{V}\mathbf{V}^H$ . Let  $\mathbf{V}_{-N}$  and  $\mathbf{V}_{-1}$  be the matrices obtained from  $\mathbf{V}$  by removing its last and first row, respectively. By the structure of  $\mathbf{T}$ , we have that  $\mathbf{V}_{-N}\mathbf{V}_{-N}^H = \mathbf{V}_{-1}\mathbf{V}_{-1}^H$ . Thus there must exist an  $r \times r$  unitary matrix  $\mathbf{Q}$  satisfying  $\mathbf{V}_{-1} = \mathbf{V}_{-N}\mathbf{Q}$  (see, e.g., [127, Theorem 7.3.11]). It follows that  $\mathbf{V}_j = \mathbf{V}_1\mathbf{Q}^{j-1}$ ,  $j = 2, \dots, N$  and therefore,

$$u_j = \mathbf{V}_1\mathbf{Q}^{1-j}\mathbf{V}_1^H, \quad j = 1, \dots, N, \quad (11.103)$$

where  $\mathbf{V}_j$  is the  $j$ th row of  $\mathbf{V}$ . Next, write the eigen-decomposition of the unitary matrix  $\mathbf{Q}$ , which is guaranteed to exist, as

$$\mathbf{Q} = \tilde{\mathbf{Q}} \text{diag}(z_1, \dots, z_r) \tilde{\mathbf{Q}}^H, \quad (11.104)$$

where  $\tilde{\mathbf{Q}}$  is also an  $r \times r$  unitary matrix and  $z_k$ 's are the eigenvalues. Since the eigenvalues of a unitary matrix must have unit magnitude, we can find  $f_k \in \mathbb{T}$ ,  $k = 1, \dots, r$  satisfying  $z_k = e^{i2\pi f_k}$ ,  $k = 1, \dots, r$ . Inserting Eq. (11.104) into Eq. (11.103) and letting  $p_k = |\mathbf{V}_1\tilde{\mathbf{Q}}_{:k}|^2 > 0$ ,  $k = 1, \dots, r$ , where  $\tilde{\mathbf{Q}}_{:k}$  denotes the  $k$ th column of  $\tilde{\mathbf{Q}}$ , we have that

$$u_j = \sum_{k=1}^r p_k e^{-i2\pi(j-1)f_k}. \quad (11.105)$$

It follows that Eq. (11.102) holds. Moreover,  $f_k, k = 1, \dots, r$  are distinct since otherwise  $\text{rank}(\mathbf{T}) < r$ , which cannot be true.

We now consider the case of  $r = N$  for which  $\mathbf{T} > 0$ . Let us arbitrarily choose  $f_N \in \mathbb{T}$  and let  $p_N = (\mathbf{a}^H(f_N)\mathbf{T}^{-1}\mathbf{a}(f_N))^{-1} > 0$ . Moreover, we define a new vector  $\mathbf{u}' \in \mathbb{C}^N$  by

$$u'_j = u_j - p_N e^{-i2\pi(j-1)f_N}. \quad (11.106)$$

It can be readily verified that

$$\mathbf{T}(\mathbf{u}') = \mathbf{T}(\mathbf{u}) - p_N \mathbf{a}(f_N) \mathbf{a}^H(f_N), \quad (11.107)$$

$$\mathbf{T}(\mathbf{u}') \geq 0, \quad (11.108)$$

$$\text{rank}(\mathbf{T}(\mathbf{u}')) = N - 1. \quad (11.109)$$

Therefore, from the result in the case of  $r \leq N - 1$  proven above,  $\mathbf{T}(\mathbf{u}')$  admits a Vandermonde decomposition as in Eq. (11.102) with  $r = N - 1$ . It then follows from Eq. (11.107) that  $\mathbf{T}(\mathbf{u})$  admits a Vandermonde decomposition with  $N$  ‘atoms.’

We finally show the uniqueness in the case of  $r \leq N - 1$ . To do so, suppose there exists another decomposition  $\mathbf{T} = \mathbf{A}(f')\mathbf{P}'\mathbf{A}^H(f')$  in which  $p'_j > 0, j = 1, \dots, r$  and  $f'_j \in \mathbb{T}$  are distinct. It follows from the equation

$$\mathbf{A}(f')\mathbf{P}'\mathbf{A}^H(f') = \mathbf{A}(f)\mathbf{P}\mathbf{A}^H(f) \quad (11.110)$$

that there exists an  $r \times r$  unitary matrix  $\mathbf{Q}'$  such that  $\mathbf{A}(f')\mathbf{P}'^{\frac{1}{2}} = \mathbf{A}(f)\mathbf{P}^{\frac{1}{2}}\mathbf{Q}'$  and therefore,

$$\mathbf{A}(f') = \mathbf{A}(f)\mathbf{P}^{\frac{1}{2}}\mathbf{Q}'\mathbf{P}'^{-\frac{1}{2}}. \quad (11.111)$$

This means that for every  $j = 1, \dots, r$ ,  $\mathbf{a}(f'_j)$  lies in the range space spanned by  $\{\mathbf{a}(f_k)\}_{k=1}^r$ . By the fact that  $r \leq N - 1$  and that any  $N$  atoms  $\mathbf{a}(f_k)$  with distinct  $f_k$ 's are linearly independent, we have that  $f'_j \in \{f_k\}_{k=1}^r$  and thus the two sets  $\{f'_j\}_{j=1}^r$  and  $\{f_k\}_{k=1}^r$  are identical. It follows that the above two decompositions of  $\mathbf{T}(\mathbf{u})$  must be identical.  $\square$

Note that the proof of Theorem 11.5 provides a computational approach to the Vandermonde decomposition. We simply consider the case of  $r \leq N - 1$ , since in the case of  $r = N$  we can arbitrarily choose  $f_N \in \mathbb{T}$  first. We use the following result:

$$(\mathbf{V}_{-N}^H \mathbf{V}_{-1} - z_k \mathbf{V}_{-N}^H \mathbf{V}_{-N}) \tilde{\mathbf{Q}}_{:k} = 0, \quad (11.112)$$

which can be shown along the lines of the above proof. To retrieve the frequencies and the powers from  $\mathbf{T}$ , we first compute  $\mathbf{V} \in \mathbb{C}^{N \times r}$  satisfying  $\mathbf{T} = \mathbf{V}\mathbf{V}^H$  using, e.g., the Cholesky decomposition. After that, we use Eq. (11.112) and compute  $z_k$  and  $\tilde{\mathbf{Q}}_{:k}$ ,  $k = 1, \dots, r$  as the eigenvalues and the normalized eigenvectors of the matrix pencil

$(\mathbf{V}_{-N}^H \mathbf{V}_{-1}, \mathbf{V}_{-N}^H \mathbf{V}_{-N})$ . Finally, we obtain  $f_k = \frac{1}{2\pi} \text{Im}(\ln z_k) \in \mathbb{T}$  and  $p_k = |\mathbf{V}_1 \tilde{\mathbf{Q}}_{:k}|^2$ ,  $k = 1, \dots, r$ , where  $\text{Im}$  gives the imaginary part of its argument. In fact, this matrix pencil approach is similar to the ESPRIT algorithm that computes the frequency estimates from an estimate of the data covariance matrix.

In the presence of homoscedastic noise, the data covariance matrix  $\mathbf{R}$  remains Toeplitz. In this case, it is natural to decompose the Toeplitz covariance matrix as the sum of the signal covariance and the noise covariance. Consequently, the following corollary of [Theorem 11.5](#) can be useful in such a case. The proof is straightforward and will be omitted.

**Corollary 11.1** Any PSD Toeplitz matrix  $\mathbf{T}(\mathbf{u}) \in \mathbb{C}^{N \times N}$  can be uniquely decomposed as:

$$\mathbf{T} = \sum_{k=1}^r p_k \mathbf{a}(f_k) \mathbf{a}^H(f_k) + \sigma \mathbf{I} = \mathbf{A}(\mathbf{f}) \text{diag}(\mathbf{p}) \mathbf{A}^H(\mathbf{f}) + \sigma \mathbf{I}, \quad (11.113)$$

where  $\sigma = \lambda_{\min}(\mathbf{T})$  (the smallest eigenvalue of  $\mathbf{T}$ ),  $r = \text{rank}(\mathbf{T} - \sigma \mathbf{I}) < N$ ,  $p_k > 0$ , and  $f_k \in \mathbb{T}$ ,  $k = 1, \dots, r$  are distinct.

*Remark 11.1.* Note that the uniqueness of the decomposition in [Corollary 11.1](#) is guaranteed by the condition that  $\sigma = \lambda_{\min}(\mathbf{T})$ . If the condition is violated by letting  $0 \leq \sigma < \lambda_{\min}(\mathbf{T})$  (in such a case  $\mathbf{T}$  has full rank and  $r \geq N$ ), then the decomposition in Eq. (11.113) cannot be unique.

The Vandermonde decomposition of Toeplitz covariance matrices forms an important tool in several recently proposed gridless sparse methods. In particular, these methods transform the frequency estimation problem into the estimation of a PSD Toeplitz matrix in which the frequencies are encoded. Once the matrix is computed, the frequencies can be retrieved from its Vandermonde decomposition. Therefore, these gridless sparse methods can be viewed as being covariance-based by interpreting the Toeplitz matrix as the data covariance matrix (though it might not be, since certain statistical assumptions may not be satisfied). In contrast to conventional subspace methods that estimate the frequencies directly from the sample covariance matrix, the gridless methods utilize more sophisticated optimization approaches to estimate the data covariance matrix by exploiting its special structures, e.g., Toeplitz, low rank and PSDness, and therefore are expected to achieve superior performance.

### 11.6.3 THE SINGLE SNAPSHOT CASE

In this subsection we introduce several gridless sparse methods for DOA/frequency estimation in the single snapshot case (a.k.a. the line spectral estimation problem). Two kinds of methods will be discussed: deterministic optimization methods, e.g., the atomic norm and the Hankel-based nuclear norm methods [83, 100, 128–138], and a covariance fitting method that is a gridless version of SPICE [84, 92, 133, 139–142]. The connections between these methods will also be investigated.

By “deterministic” we mean that we do not make any statistical assumptions on the signal of interest. Instead, the signal is deterministic and it is sought as the sparsest candidate, measured by a certain sparse metric, among a prescribed set.

### 11.6.3.1 A general framework for deterministic methods

In the single snapshot case, corresponding to Eq. (11.99), the data model in the SLA case is given by:

$$\mathbf{y}_\Omega = \mathbf{z}_\Omega + \mathbf{e}_\Omega, \quad \mathbf{z} = \mathbf{A}(f)\mathbf{s}, \quad (11.114)$$

where  $\mathbf{z}$  denotes the noiseless signal. Note that the ULA is a special case with  $\Omega = \{1, \dots, N\}$ . For deterministic sparse methods, in general, we need to solve a constrained optimization problem of the following form:

$$\min_z \mathcal{M}(z), \text{ subject to } \|z_\Omega - \mathbf{y}_\Omega\|_2 \leq \eta, \quad (11.115)$$

where the noise is assumed to be bounded:  $\|\mathbf{e}_\Omega\|_2 \leq \eta$ . In Eq. (11.115),  $\mathbf{z}$  is the sinusoidal signal of interest, and  $\mathcal{M}(z)$  denotes a sparse metric that is defined such that by minimizing  $\mathcal{M}(z)$  the number of components/atoms  $\mathbf{a}(f)$  used to express  $\mathbf{z}$  is reduced, and these atoms give the frequency estimates. Instead of Eq. (11.115), we may solve a regularized optimization problem given by:

$$\min_z \lambda \mathcal{M}(z) + \frac{1}{2} \|z_\Omega - \mathbf{y}_\Omega\|_2^2, \quad (11.116)$$

where the noise is typically assumed to be Gaussian and  $\lambda > 0$  is a regularization parameter. In the extreme noiseless case, by letting  $\eta \rightarrow 0$  and  $\lambda \rightarrow 0$ , both Eqs. (11.115) and (11.116) reduce to the following problem:

$$\min_z \mathcal{M}(z), \text{ subject to } z_\Omega = \mathbf{y}_\Omega. \quad (11.117)$$

We next discuss different choices of  $\mathcal{M}(z)$ .

### 11.6.3.2 Atomic $\ell_0$ norm

To promote sparsity to the greatest extent possible, inspired by the literature on sparse recovery and compressed sensing, the natural choice of  $\mathcal{M}(z)$  is an  $\ell_0$  norm like sparse metric, referred to as the atomic  $\ell_0$  (pseudo-)norm. Let us formally define the set of atoms used here:

$$\mathcal{A} = \{\mathbf{a}(f, \phi) = \mathbf{a}(f)\phi : f \in \mathbb{T}, \phi \in \mathbb{C}, |\phi| = 1\}. \quad (11.118)$$

It is evident from Eq. (11.114) that the true signal  $\mathbf{z}$  is a linear combination of  $K$  atoms in the atomic set  $\mathcal{A}$ . The atomic  $\ell_0$  (pseudo-)norm, denoted by  $\|\mathbf{z}\|_{\mathcal{A}, 0}$ , is defined as the minimum number of atoms in  $\mathcal{A}$  that can synthesize  $\mathbf{z}$ :

$$\begin{aligned} \|\mathbf{z}\|_{\mathcal{A}, 0} &= \inf_{c_k, f_k, \phi_k} \left\{ \mathcal{K} : \mathbf{z} = \sum_{k=1}^{\mathcal{K}} \mathbf{a}(f_k, \phi_k) c_k, f_k \in \mathbb{T}, |\phi_k| = 1, c_k > 0 \right\} \\ &= \inf_{f_k, s_k} \left\{ \mathcal{K} : \mathbf{z} = \sum_{k=1}^{\mathcal{K}} \mathbf{a}(f_k) s_k, f_k \in \mathbb{T} \right\}. \end{aligned} \quad (11.119)$$

To provide a finite-dimensional formulation for  $\|z\|_{\mathcal{A},0}$ , the Vandermonde decomposition of Toeplitz covariance matrices is invoked. To be specific, let  $\mathbf{T}(\mathbf{u})$  be a Toeplitz matrix and impose the condition that

$$\begin{bmatrix} x & z^H \\ z & \mathbf{T}(\mathbf{u}) \end{bmatrix} \geq 0, \quad (11.120)$$

where  $x$  is a free variable to be optimized. It follows from Eq. (11.120) that  $\mathbf{T}(\mathbf{u})$  is PSD and thus admits a  $\text{rank}(\mathbf{T}(\mathbf{u}))$ -atomic Vandermonde decomposition. Moreover,  $z$  lies in the range space of  $\mathbf{T}(\mathbf{u})$ . Therefore,  $z$  can be expressed by  $\text{rank}(\mathbf{T}(\mathbf{u}))$  atoms. This means that the atomic  $\ell_0$  norm is linked to the rank of  $\mathbf{T}(\mathbf{u})$ . Formally, we have the following result.

**Theorem 11.6 ([130])**  $\|z\|_{\mathcal{A},0}$  defined in Eq. (11.119) equals the optimal value of the following rank minimization problem:

$$\min_{x, \mathbf{u}} \text{rank}(\mathbf{T}(\mathbf{u})), \text{ subject to Eq. (11.120).} \quad (11.121)$$

By Theorem 11.6, the atomic  $\ell_0$  norm method needs to solve a rank minimization problem that, as might have been expected, cannot be easily solved. By the rank minimization formulation, the frequencies of interest are actually encoded in the PSD Toeplitz matrix  $\mathbf{T}(\mathbf{u})$ . If  $\mathbf{T}(\mathbf{u})$  can be solved for, then the frequencies can be retrieved from its Vandermonde decomposition. Therefore, the Toeplitz matrix  $\mathbf{T}(\mathbf{u})$  in Eq. (11.121) can be viewed as the covariance matrix of the noiseless signal  $z$  as if certain statistical assumptions were satisfied (however, those assumptions are not required here). Note that the Toeplitz structure of the covariance matrix is explicitly enforced, the PSDness is imposed by the constraint in Eq. (11.120), and the low-rankness is the objective.

### 11.6.3.3 Atomic norm

A practical choice of the sparse metric  $\mathcal{M}(z)$  is the atomic norm that is a convex relaxation of the atomic  $\ell_0$  norm. The resulting optimization problems in Eqs. (11.115)–(11.117) are referred to as atomic norm minimization (ANM). The concept of atomic norm was first proposed in [143] and it generalizes several norms commonly used for sparse representation and recovery, e.g., the  $\ell_1$  norm and the nuclear norm, for appropriately chosen atoms. The atomic norm is basically equivalent to the total variation norm [144] that was adopted, e.g., in [100]. We have decided to use the atomic norm in this article since it is simpler to present and easier to understand. Formally, the atomic norm is defined as the gauge function of  $\text{conv}(\mathcal{A})$ , the convex hull of  $\mathcal{A}$  [143]:

$$\begin{aligned}
\|z\|_{\mathcal{A}} &= \inf\{t > 0 : z \in t\text{conv}(\mathcal{A})\} \\
&= \inf_{c_k, f_k, \phi_k} \left\{ \sum_k c_k : z = \sum_k \mathbf{a}(f_k, \phi_k) c_k, f_k \in \mathbb{T}, |\phi_k| = 1, c_k > 0 \right\} \\
&= \inf_{f_k, s_k} \left\{ \sum_k |s_k| : z = \sum_k \mathbf{a}(f_k) s_k, f_k \in \mathbb{T} \right\}.
\end{aligned} \tag{11.122}$$

By definition, the atomic norm can be viewed as a continuous counterpart of the  $\ell_1$  norm used in the discrete setting. Different from the  $\ell_1$  norm, however, it is unclear how to compute the atomic norm from the definition. In fact, initially this has been a major obstacle in applying the atomic norm technique [143, 145]. To solve this problem, a computationally efficient SDP formulation of  $\|z\|_{\mathcal{A}}$  is provided in the following result. A proof of the result is also provided, which helps illustrate how the frequencies can be obtained.

**Theorem 11.7 ([130])**  $\|z\|_{\mathcal{A}}$  defined in Eq. (11.122) equals the optimal value of the following SDP:

$$\min_{x, u} \frac{1}{2}x + \frac{1}{2}u_1, \text{ subject to Eq. (11.120).} \tag{11.123}$$

*Proof.* Let  $F$  be the optimal value of the objective in Eq. (11.123). We need to show that  $F = \|z\|_{\mathcal{A}}$ .

We first show that  $F \leq \|z\|_{\mathcal{A}}$ . To do so, let  $z = \sum_k c_k \mathbf{a}(f_k, \phi_k) = \sum_k \mathbf{a}(f_k) s_k$  be an atomic decomposition of  $z$ . Then let  $u$  be such that  $T(u) = \sum_k c_k \mathbf{a}(f_k) \mathbf{a}^H(f_k)$  and  $x = \sum_k c_k$ . It follows that

$$\begin{bmatrix} x & z^H \\ z & T \end{bmatrix} = \sum_k c_k \begin{bmatrix} \phi_k^* \\ \mathbf{a}(f_k) \end{bmatrix} \begin{bmatrix} \phi_k^* \\ \mathbf{a}(f_k) \end{bmatrix}^H \geq 0. \tag{11.124}$$

Therefore,  $x$  and  $u$  constructed as above form a feasible solution to the problem in Eq. (11.123), at which the objective value equals

$$\frac{1}{2}x + \frac{1}{2}u_1 = \sum_k c_k. \tag{11.125}$$

It follows that  $F \leq \sum_k c_k$ . Since the inequality holds for any atomic decomposition of  $z$ , we have that  $F \leq \|z\|_{\mathcal{A}}$  by the definition of the atomic norm.

On the other hand, suppose that  $(\hat{x}, \hat{u})$  is an optimal solution to the problem in Eq. (11.123). By the fact that  $T(\hat{u}) \geq 0$  and applying Theorem 11.5, we have that  $T(\hat{u})$  admits a Vandermonde decomposition as in Eq. (11.102) with  $(r, p_k, f_k)$  denoted by  $(\hat{r}, \hat{p}_k, \hat{f}_k)$ . Moreover, since  $\begin{bmatrix} \hat{x} & z^H \\ z & T(\hat{u}) \end{bmatrix} \geq 0$ , we have that  $z$  lies in the range space of  $T(\hat{u})$  and thus has the following atomic decomposition:

$$z = \sum_{k=1}^{\hat{r}} \hat{c}_k \mathbf{a}(\hat{f}_k, \hat{\phi}_k) = \sum_{k=1}^{\hat{r}} \mathbf{a}(\hat{f}_k) \hat{s}_k. \tag{11.126}$$

Moreover, it holds that

$$\hat{x} \geq z^H [\mathbf{T}(\hat{\mathbf{u}})]^\dagger z = \sum_{k=1}^{\hat{r}} \frac{\hat{c}_k^2}{\hat{p}_k}, \quad (11.127)$$

$$\hat{u}_0 = \sum_{k=1}^{\hat{r}} \hat{p}_k. \quad (11.128)$$

It therefore follows that

$$\begin{aligned} F &= \frac{1}{2} \hat{x} + \frac{1}{2} \hat{u}_1 \\ &\geq \frac{1}{2} \sum_k \frac{\hat{c}_k^2}{\hat{p}_k} + \frac{1}{2} \sum_k \hat{p}_k \\ &\geq \sum_k \hat{c}_k \\ &\geq \|z\|_{\mathcal{A}}. \end{aligned} \quad (11.129)$$

Combining Eq. (11.129) and the inequality  $F \leq \|z\|_{\mathcal{A}}$  that was shown previously, we conclude that  $F = \|z\|_{\mathcal{A}}$ , which completes the proof. It is worth noting that by Eq. (11.129) we must have that  $\hat{p}_k = \hat{c}_k = |\hat{s}_k|$  and  $\|z\|_{\mathcal{A}} = \sum_k \hat{c}_k = \sum_k |\hat{s}_k|$ . Therefore, the atomic decomposition in Eq. (11.126) achieves the atomic norm.  $\square$

Interestingly (but not surprisingly), the SDP in Eq. (11.123) is actually a convex relaxation of the rank minimization problem in Eq. (11.121). Concretely, the second term  $\frac{1}{2} u_1$  in the objective function in Eq. (11.123) is actually the nuclear norm or the trace norm of  $\mathbf{T}(\mathbf{u})$  (up to a scaling factor), which is a commonly used convex relaxation of the matrix rank, while the first term  $\frac{1}{2} x$  is a regularization term that prevents a trivial solution.

Similar to the atomic  $\ell_0$  norm, the frequencies in the atomic norm approach are also encoded in the Toeplitz matrix  $\mathbf{T}(\mathbf{u})$ . Once the resulting SDP problem is solved, the frequencies can be retrieved from the Vandermonde decomposition of  $\mathbf{T}(\mathbf{u})$ . Therefore, similar to the  $\ell_0$  norm, the atomic norm can also be viewed as being covariance-based. The only difference lies in enforcing the low-rankness of the “covariance” matrix  $\mathbf{T}(\mathbf{u})$ . The atomic  $\ell_0$  norm directly uses the rank function that exploits the low-rankness to the greatest extent possible but cannot be practically solved. In contrast to this, the atomic norm uses a convex relaxation, the nuclear norm (or the trace norm), and provides a practically feasible approach.

In the absence of noise, the theoretical performance of ANM has been studied in [100, 130]. In the case of ULA where all the entries of  $\mathbf{y}$  are observed, the ANM problem derived from Eq. (11.117) actually admits a trivial solution  $\mathbf{z} = \mathbf{y}$ . But the following SDP resulting from Eq. (11.123) still makes sense and can be used for frequency estimation:

$$\min_{x, \mathbf{u}} \frac{1}{2} x + \frac{1}{2} u_1, \text{ subject to } \begin{bmatrix} x & \mathbf{y}^H \\ \mathbf{y} & \mathbf{T}(\mathbf{u}) \end{bmatrix} \geq \mathbf{0}. \quad (11.130)$$

Let  $\mathcal{T} = \{f_1, \dots, f_K\}$  and define the minimum separation of  $\mathcal{T}$  as the closest wrap-around distance between any two elements:

$$\Delta_{\mathcal{T}} = \inf_{1 \leq j \neq k \leq K} \min \{|f_j - f_k|, 1 - |f_j - f_k|\}. \quad (11.131)$$

The following theoretical guarantee for the atomic norm is provided in [100].

**Theorem 11.8 ([100])**  $\mathbf{y} = \sum_{j=1}^K c_j \mathbf{a}(f_j, \phi_j)$  is the unique atomic decomposition satisfying  $\|\mathbf{y}\|_{\mathcal{A}} = \sum_{j=1}^K c_j$  if  $\Delta_{\mathcal{T}} \geq \frac{1}{\lfloor (N-1)/4 \rfloor}$  and  $N \geq 257$ .

By Theorem 11.8, in the noiseless case the frequencies can be exactly recovered by solving the SDP in Eq. (11.130) if the frequencies are separated by at least  $\frac{4}{N}$  (note that this frequency separation condition is sufficient but not necessary and it has recently been relaxed to  $\frac{2.52}{N}$  in [136]). Moreover, the condition  $N \geq 257$  is a technical requirement that should not pose any serious problem in practice.

In the SLA case, the SDP resulting from Eq. (11.117) is given by:

$$\min_{x, u, z} \frac{1}{2}x + \frac{1}{2}u_1, \text{ subject to } \begin{bmatrix} x & z^H \\ z & \mathbf{T}(u) \end{bmatrix} \geq \mathbf{0}, z_{\Omega} = \mathbf{y}_{\Omega}. \quad (11.132)$$

The following result shows that the frequencies can be exactly recovered by solving Eq. (11.132) if sufficiently many samples are observed and the same frequency separation condition as above is satisfied.

**Theorem 11.9 ([130])** Suppose we observe  $\mathbf{y} = \sum_{j=1}^K c_j \mathbf{a}(f_j, \phi_j)$  on the index set  $\Omega$ , where  $\Omega \subset \{1, \dots, N\}$  is of size  $M$  and is selected uniformly at random. Assume that  $\{\phi_j\}_{j=1}^K$  are drawn i.i.d. from the uniform distribution on the complex unit circle.<sup>1</sup>

If  $\Delta_{\mathcal{T}} \geq \frac{1}{\lfloor (N-1)/4 \rfloor}$ , then there exists a numerical constant  $C$  such that

$$M \geq C \max \left\{ \log^2 \frac{N}{\delta}, K \log \frac{K}{\delta} \log \frac{N}{\delta} \right\} \quad (11.133)$$

is sufficient to guarantee that, with probability at least  $1 - \delta$ ,  $\mathbf{y}$  is the unique optimizer for Eq. (11.132) and  $\mathbf{y} = \sum_{j=1}^K c_j \mathbf{a}(f_j, \phi_j)$  is the unique atomic decomposition satisfying  $\|\mathbf{y}\|_{\mathcal{A}} = \sum_{j=1}^K c_j$ .

In the presence of noise, the SDP resulting from the unconstrained formulation in Eq. (11.116) is given by:

$$\min_{x, u, z} \frac{\lambda}{2}(x + u_1) + \frac{1}{2}\|z_{\Omega} - \mathbf{y}_{\Omega}\|_2^2, \text{ subject to } \begin{bmatrix} x & z^H \\ z & \mathbf{T}(u) \end{bmatrix} \geq \mathbf{0}. \quad (11.134)$$

---

<sup>1</sup>This condition has been relaxed in [22], where it was assumed that  $\{\phi_j\}_{j=1}^K$  are independent with zero mean.

While it is clear that the regularization parameter  $\lambda$  is used to balance the signal sparsity and the data fidelity, it is less clear how to choose it. Under the assumption of i.i.d. Gaussian noise, this choice has been studied in [83, 131, 133]. For ULAs the paper [83] shows that if we let  $\lambda \approx \sqrt{M \log M \sigma}$ , where  $\sigma$  denotes the noise variance, then the signal estimate  $\hat{\mathbf{z}}$  given by Eq. (11.134) has the following per-element expected reconstruction error:

$$\frac{1}{M} \mathbb{E} \|\hat{\mathbf{z}} - \mathbf{z}\|_2^2 \leq \sqrt{\frac{\log M}{M}} \sigma \cdot \sum_{k=1}^K c_k. \quad (11.135)$$

This error bound implies that if  $K = o\left(\sqrt{\frac{M}{\log M}}\right)$ , then the estimate  $\hat{\mathbf{z}}$  is statistically

consistent. Moreover, the paper [131] shows that if we let  $\lambda = C\sqrt{M \log M \sigma}$ , where  $C > 1$  is a constant (not explicitly given), and if the frequencies are sufficiently separated as in [Theorem 11.8](#), then the following error bound can be obtained with high probability:

$$\frac{1}{M} \|\hat{\mathbf{z}} - \mathbf{z}\|_2^2 = O\left(\frac{K \log M}{M} \sigma\right), \quad (11.136)$$

which is nearly minimax optimal. This implies that the estimate is consistent if  $K = o\left(\frac{M}{\log M}\right)$ . Furthermore, the frequencies and the amplitudes can be stably estimated as well [131]. Finally, note that the result in [83] has been generalized to the SLA case in [133]. It was shown that if we let  $\lambda \approx \sqrt{M \log N \sigma}$  in such a case, it holds similarly to Eq. (11.135) that

$$\frac{1}{M} \mathbb{E} \|\hat{\mathbf{z}}_\Omega - \mathbf{z}_\Omega\|_2^2 \leq \sqrt{\frac{\log N}{M}} \sigma \cdot \sum_{k=1}^K c_k. \quad (11.137)$$

This means that the estimate  $\hat{\mathbf{z}}_\Omega$  is consistent if  $K = o\left(\sqrt{\frac{M}{\log N}}\right)$ .

#### 11.6.3.4 Hankel-based nuclear norm

Another choice of  $\mathcal{M}(\mathbf{z})$  is the Hankel-based nuclear norm that was proposed in [132]. This metric is introduced based on the following observation. Given  $\mathbf{z}$  as in Eq. (11.114), let us form the Hankel matrix:

$$\mathbf{H}(\mathbf{z}) = \begin{bmatrix} z_1 & z_2 & \dots & z_n \\ z_2 & z_3 & \dots & z_{n+1} \\ \vdots & \vdots & \ddots & \vdots \\ z_m & z_{m+1} & \dots & z_N \end{bmatrix}, \quad (11.138)$$

where  $m + n = N + 1$ . It follows that

$$\mathbf{H}(\mathbf{z}) = \sum_{k=1}^K s_k \begin{bmatrix} 1 \\ e^{i2\pi f_k} \\ \vdots \\ e^{i2\pi(m-1)f_k} \end{bmatrix} \begin{bmatrix} 1 & e^{i2\pi f_k} & \dots & e^{i2\pi(n-1)f_k} \end{bmatrix}. \quad (11.139)$$

If  $K < \min(m, n)$ , then we have that  $\mathbf{H}(z)$  is a low rank matrix with

$$\text{rank}(\mathbf{H}(z)) = K. \quad (11.140)$$

To reconstruct  $z$ , therefore, we may consider the reconstruction of  $\mathbf{H}(z)$  by choosing the sparse metric as  $\text{rank}(\mathbf{H}(z))$ . If  $z$  can be determined for the resulting rank minimization problem, then the frequencies may be recovered from  $z$ .

Since the rank minimization cannot be easily solved, we seek a convex relaxation of  $\text{rank}(\mathbf{H}(z))$ . The nuclear norm is a natural choice, which leads to:

$$\mathcal{M}(z) = \|\mathbf{H}(z)\|_{\star}. \quad (11.141)$$

The optimization problems resulting from Eqs. (11.115)–(11.117) by using Eq. (11.141) are referred to as enhanced matrix completion (EMaC) in [132]. Note that the nuclear norm can be formulated as the following SDP [146]:

$$\|\mathbf{H}(z)\|_{\star} = \min_{Q_1, Q_2} \frac{1}{2} [\text{Tr}(Q_1) + \text{Tr}(Q_2)], \text{ subject to } \begin{bmatrix} Q_1 & \mathbf{H}(z)^H \\ \mathbf{H}(z) & Q_2 \end{bmatrix} \geq 0. \quad (11.142)$$

As a result, like the atomic norm method, the EMaC problems can be cast as SDP and solved using off-the-shelf solvers.

Theoretical guarantees for EMaC have been provided in the SLA case in [132] which, to some extent, are similar to those for the atomic norm method. In particular, it was shown that the signal  $z$  can be exactly recovered in the absence of noise and stably recovered in the presence of bounded noise if the number of measurements  $M$  exceeds a constant times the number of sinusoids  $K$  up to a polylog factor, as given by Eq. (11.133), and if a certain coherence condition is satisfied. It is argued in [132] that the coherence condition required by EMaC can be weaker than the frequency separation condition required by ANM and thus higher resolution might be obtained by EMaC as compared to ANM. Connections between the two methods will be studied in the following subsection.

### 11.6.3.5 Connection between ANM and EMaC

To investigate the connection between ANM and EMaC, we define the following set of complex exponentials as a new atomic set:

$$\mathcal{A}' = \left\{ \mathbf{a}'(\phi) = [1, \phi, \dots, \phi^{N-1}]^T : \phi \in \mathbb{C} \right\}. \quad (11.143)$$

It is evident that, as compared to  $\mathcal{A}$ , the complex exponentials are restricted to have constant modulus in  $\mathcal{A}$  with  $|\phi| = 1$  and thus  $\mathcal{A}$  is a subset of  $\mathcal{A}'$ . For any  $z \in \mathbb{C}^N$ , we can similarly define the atomic  $\ell_0$  norm with respect to  $\mathcal{A}'$ , denoted by  $\|z\|_{\mathcal{A}', 0}$ . We have the following result.

**Theorem 11.10** *For any  $z$  it holds that*

$$\|z\|_{\mathcal{A}', 0} \geq \text{rank}(\mathbf{H}(z)). \quad (11.144)$$

*Moreover, if  $\text{rank}(\mathbf{H}(z)) < \min(m, n)$ , then*

$$\|z\|_{\mathcal{A}', 0} = \text{rank}(\mathbf{H}(z)) \quad (11.145)$$

*except for degenerate cases.*

*Proof.* Suppose that  $\|z\|_{\mathcal{A}',0} = K'$ . This means that there exists a  $K'$ -atomic decomposition for  $z$  with respect to  $\mathcal{A}'$ :

$$z = \sum_{k=1}^{K'} \alpha(\phi_k) s'_k, \quad \phi_k \in \mathbb{C}. \quad (11.146)$$

It follows that  $\mathbf{H}(z)$  admits a decomposition similar to Eq. (11.139) and therefore that  $\text{rank}(\mathbf{H}(z)) \leq K'$  and hence Eq. (11.144) holds.

The second part can be shown by applying the Kronecker's theorem for Hankel matrices (see, e.g., [147]). In particular, the Kronecker's theorem states that if  $\text{rank}(\mathbf{H}(z)) = K' < \min(m, n)$ , then  $z$  can be written as in Eq. (11.146) except for degenerate cases. According to the definition of  $\|z\|_{\mathcal{A}',0}$ , we have that

$$\|z\|_{\mathcal{A}',0} \leq K' = \text{rank}(\mathbf{H}(z)). \quad (11.147)$$

This together with Eq. (11.144) implies Eq. (11.145).  $\square$

By Theorem 11.10, we have linked  $\text{rank}(\mathbf{H}(z))$ , which motivated the use of its convex relaxation  $\|\mathbf{H}(z)\|_*$ , to an atomic  $\ell_0$  norm. In the regime of interest here  $\mathbf{H}(z)$  is low-rank and hence  $\text{rank}(\mathbf{H}(z))$  is almost identical to the atomic  $\ell_0$  norm induced by  $\mathcal{A}'$ . To compare  $\|z\|_{\mathcal{A},0}$  and  $\|z\|_{\mathcal{A}',0}$ , we have the following result.

**Theorem 11.11** *For any  $z$  it holds that*

$$\|z\|_{\mathcal{A}',0} \leq \|z\|_{\mathcal{A},0}. \quad (11.148)$$

*Proof.* The inequality is a direct consequence of the fact that  $\mathcal{A} \subset \mathcal{A}'$ .  $\square$

By Theorem 11.11 the newly defined  $\|z\|_{\mathcal{A}',0}$ , which is closely related to  $\text{rank}(\mathbf{H}(z))$ , is actually a lower bound on  $\|z\|_{\mathcal{A},0}$ . It is worth noting that this lower bound is obtained by ignoring a known structure of the signal: from  $\mathcal{A}$  to  $\mathcal{A}'$  we have neglected the prior knowledge that each exponential component of  $z$  has constant modulus. As a consequence, using  $\|z\|_{\mathcal{A}',0}$  as the sparse metric instead of  $\|z\|_{\mathcal{A},0}$ , we cannot guarantee in general, especially in the noisy case, that each component of the obtained signal  $z$  corresponds to one frequency. Note that this is also true for the convex relaxation metric  $\|\mathbf{H}(z)\|_*$  as compared to  $\|z\|_{\mathcal{A}}$ . In contrast to this, the frequencies can be directly retrieved from the solution of the atomic norm method. From this point of view, the atomic norm method may be expected to outperform EMaC due to its better capability to capture the signal structure.

On the other hand, EMaC might have higher resolution than the atomic norm method. This is indeed true in the noiseless ULA case where the signal  $z$  is completely known. In this extreme case EMaC does not suffer from any resolution limit, while the atomic norm requires a frequency separation condition for its successful operation (at least theoretically).

### 11.6.3.6 Covariance fitting method: Gridless SPICE (GLS)

GLS was introduced in [92, 133] as a gridless version of the SPICE method presented in Section 11.4.5. Since SPICE is covariance-based and the data covariance matrix is a highly nonlinear function of the DOA parameters of interest, gridding is performed in SPICE to linearize the problem based on the zeroth order approximation. But this is not required in the case of ULAs or SLAs. The key idea of GLS is to re-parameterize the data covariance matrix using a PSD Toeplitz matrix  $\mathbf{T}(\mathbf{u})$ , which is linear in the new parameter vector  $\mathbf{u}$ , by making use of the Vandermonde decomposition of Toeplitz covariance matrices. To derive GLS, naturally, we make the same assumptions as for SPICE.

We first consider the ULA case. Assume that the noise is homoscedastic (note that, like SPICE, GLS can be extended to the case of heteroscedastic noise). It follows from the arguments in Section 11.6.2 that the data covariance matrix  $\mathbf{R}$  is a Toeplitz matrix. Therefore,  $\mathbf{R}$  can be linearly re-parameterized as:

$$\mathbf{R} = \mathbf{T}(\mathbf{u}), \quad \mathbf{T}(\mathbf{u}) \geq 0. \quad (11.149)$$

For a single snapshot, SPICE minimizes the following covariance fitting criterion:

$$\left\| \mathbf{R}^{-\frac{1}{2}}(\mathbf{y}\mathbf{y}^H - \mathbf{R}) \right\|_F^2 = \|\mathbf{y}\|_2^2 \cdot \mathbf{y}^H \mathbf{R}^{-1} \mathbf{y} + \text{Tr}(\mathbf{R}) - 2\|\mathbf{y}\|_2^2. \quad (11.150)$$

Inserting Eq. (11.149) into Eq. (11.150), the resulting GLS optimization problem is given by:

$$\begin{aligned} & \min_{\mathbf{u}} \|\mathbf{y}\|_2^2 \cdot \mathbf{y}^H \mathbf{T}^{-1}(\mathbf{u}) \mathbf{y} + \text{Tr}(\mathbf{T}(\mathbf{u})), \text{ subject to } \mathbf{T}(\mathbf{u}) \geq 0 \\ & \Leftrightarrow \min_{x, \mathbf{u}} \|\mathbf{y}\|_2^2 x + M u_1, \text{ subject to } \mathbf{T}(\mathbf{u}) \geq \mathbf{0} \text{ and } x \geq \mathbf{y}^H \mathbf{T}^{-1}(\mathbf{u}) \mathbf{y} \\ & \Leftrightarrow \min_{x, \mathbf{u}} \|\mathbf{y}\|_2^2 x + M u_1, \text{ subject to } \begin{bmatrix} x & \mathbf{y}^H \\ \mathbf{y} & \mathbf{T}(\mathbf{u}) \end{bmatrix} \geq \mathbf{0}. \end{aligned} \quad (11.151)$$

Therefore, the covariance fitting problem has been cast as an SDP that can be solved in a polynomial time. Once the problem is solved, the data covariance estimate  $\hat{\mathbf{R}} = \mathbf{T}(\hat{\mathbf{u}})$  is obtained, where  $\hat{\mathbf{u}}$  denotes the solution of  $\mathbf{u}$ . Finally, the estimates of the parameters  $(\hat{\mathbf{f}}, \hat{\mathbf{p}}, \hat{\sigma})$  can be obtained from the decomposition of  $\hat{\mathbf{R}}$  by applying Corollary 11.1. Moreover, we note that the GLS optimization problem in Eq. (11.151) is very similar to the SDP for the atomic norm. Their connection will be discussed later in more detail.

In the SLA case, corresponding to Eq. (11.149), the covariance fitting criterion of SPICE is given by:

$$\left\| \mathbf{R}_{\Omega}^{-\frac{1}{2}}(\mathbf{y}_{\Omega} \mathbf{y}_{\Omega}^H - \mathbf{R}_{\Omega}) \right\|_F^2 = \|\mathbf{y}_{\Omega}\|_2^2 \cdot \mathbf{y}_{\Omega}^H \mathbf{R}_{\Omega}^{-1} \mathbf{y}_{\Omega} + \text{Tr}(\mathbf{R}_{\Omega}) - 2\|\mathbf{y}_{\Omega}\|_2^2. \quad (11.152)$$

where  $\mathbf{R}_\Omega$  denotes the covariance matrix of  $\mathbf{y}_\Omega$ . To explicitly express  $\mathbf{R}_\Omega$ , we let  $\mathbf{\Gamma}_\Omega \in \{0,1\}^{M \times N}$  be the row-selection matrix satisfying

$$\mathbf{y}_\Omega = \mathbf{\Gamma}_\Omega \mathbf{y}, \quad (11.153)$$

where  $\mathbf{y} \in \mathbb{C}^N$  denotes the full data vector of the virtual  $N$ -element ULA. More concretely,  $\mathbf{\Gamma}_\Omega$  is such that its  $j$ th row contains all 0s but a single 1 at the  $\Omega_j$ th position. It follows that

$$\mathbf{R}_\Omega = \mathbb{E} \mathbf{y}_\Omega \mathbf{y}_\Omega^H = \mathbf{\Gamma}_\Omega \cdot \mathbb{E} \mathbf{y} \mathbf{y}^H \cdot \mathbf{\Gamma}_\Omega^T = \mathbf{\Gamma}_\Omega \mathbf{R} \mathbf{\Gamma}_\Omega^T. \quad (11.154)$$

This means that  $\mathbf{R}_\Omega$  is a submatrix of the covariance matrix  $\mathbf{R}$  of the virtual full data  $\mathbf{y}$ . Therefore, using the parameterization of  $\mathbf{R}$  as in Eq. (11.149),  $\mathbf{R}_\Omega$  can be linearly parameterized as:

$$\mathbf{R}_\Omega = \mathbf{\Gamma}_\Omega \mathbf{T}(\mathbf{u}) \mathbf{\Gamma}_\Omega^T, \quad \mathbf{T}(\mathbf{u}) \geq 0. \quad (11.155)$$

Inserting Eq. (11.155) into Eq. (11.152), we have that the GLS optimization problem resulting from Eq. (11.152) can be cast as the following SDP:

$$\min_{x, \mathbf{u}} \|\mathbf{y}_\Omega\|_2^2 x + M u_1, \text{ subject to } \begin{bmatrix} x & \mathbf{y}_\Omega^H \\ \mathbf{y}_\Omega & \mathbf{\Gamma}_\Omega \mathbf{T}(\mathbf{u}) \mathbf{\Gamma}_\Omega^T \end{bmatrix} \geq 0. \quad (11.156)$$

Once the SDP is solved, the parameters of interest can be retrieved from the full data covariance estimate  $\hat{\mathbf{R}} = \mathbf{T}(\hat{\mathbf{u}})$  as in the ULA case.

GLS is guaranteed to produce a sparse solution with at most  $N - 1$  sources. This is a direct consequence of the frequency retrieval step, see Corollary 11.1. In general, GLS overestimates the true source number  $K$  in the presence of noise. This is reasonable because in GLS we do not assume any knowledge of the source number or of the noise variance(s).

An automatic source number estimation approach (a.k.a. model order selection) has been proposed in [133] based on the eigenvalues of the data covariance estimate  $\hat{\mathbf{R}}$ . The basic intuition behind it is that the larger eigenvalues correspond to the sources while the smaller ones are caused more likely by noise. The SORTE algorithm [148] is adopted in [133] to identify these two groups of eigenvalues and thus provide an estimate of the source number. Furthermore, based on the source number, the frequency estimates can be refined by using a subspace method such as MUSIC. Readers are referred to [133] for details.

### 11.6.3.7 Connection between ANM and GLS

GLS is strongly connected to ANM. In this subsection, we show that GLS is equivalent to ANM as if there were no noise in  $\mathbf{y}$  and with a slightly different frequency retrieval process [133]. We note that a similar connection in the discrete setting has been provided in [90, 91].

In the ULA case this connection can be shown based on the following equivalences/equalities:

$$\begin{aligned}
 \text{Eq. (11.151)} &\Leftrightarrow \min_{\mathbf{u}} \|\mathbf{y}\|_2^2 \cdot \mathbf{y}^H \mathbf{T}^{-1}(\mathbf{u}) \mathbf{y} + M u_1, \text{ subject to } \mathbf{T}(\mathbf{u}) \geq \mathbf{0} \\
 &\Leftrightarrow \min_{\mathbf{u}} M \left\{ \left[ \frac{\|\mathbf{y}\|_2}{\sqrt{M}} \mathbf{y}^H \right] \mathbf{T}^{-1}(\mathbf{u}) \left[ \frac{\|\mathbf{y}\|_2}{\sqrt{M}} \mathbf{y} \right] + u_1 \right\}, \text{ subject to } \mathbf{T}(\mathbf{u}) \geq \mathbf{0} \\
 &\Leftrightarrow \min_{x, \mathbf{u}} M \{x + u_1\}, \text{ subject to } \begin{bmatrix} x & \frac{\|\mathbf{y}\|_2}{\sqrt{M}} \mathbf{y}^H \\ \frac{\|\mathbf{y}\|_2}{\sqrt{M}} \mathbf{y} & \mathbf{T}(\mathbf{u}) \end{bmatrix} \geq \mathbf{0} \\
 &\Leftrightarrow 2M \left\| \frac{\|\mathbf{y}\|_2}{\sqrt{M}} \mathbf{y} \right\|_{\mathcal{A}} \\
 &\Leftrightarrow 2\sqrt{M} \|\mathbf{y}\|_2 \cdot \|\mathbf{y}\|_{\mathcal{A}}.
 \end{aligned} \tag{11.157}$$

This means that GLS is equivalent to computing the atomic norm of the noisy data  $\mathbf{y}$  (up to a scaling factor).

In the SLA case, we need the following equality that holds for  $\mathbf{R} > \mathbf{0}$  [133]:

$$\mathbf{y}_{\Omega}^H [\Gamma_{\Omega} \mathbf{R} \Gamma_{\Omega}^T]^{-1} \mathbf{y}_{\Omega} = \min_z z^H \mathbf{R}^{-1} z, \text{ subject to } z_{\Omega} = \mathbf{y}_{\Omega}. \tag{11.158}$$

As a result, we have that

$$\begin{aligned}
 \text{Eq. (11.156)} &\Leftrightarrow \min_{\mathbf{u}} \|\mathbf{y}_{\Omega}\|_2^2 \cdot \mathbf{y}_{\Omega}^H [\Gamma_{\Omega} \mathbf{T}(\mathbf{u}) \Gamma_{\Omega}^T]^{-1} \mathbf{y}_{\Omega} + M u_1, \text{ subject to } \mathbf{T}(\mathbf{u}) \geq \mathbf{0} \\
 &\Leftrightarrow \min_{\mathbf{u}, z} \|\mathbf{y}_{\Omega}\|_2^2 \cdot z^H \mathbf{T}^{-1}(\mathbf{u}) z + M u_1, \text{ subject to } \mathbf{T}(\mathbf{u}) \geq \mathbf{0}, z_{\Omega} = \mathbf{y}_{\Omega} \\
 &\Leftrightarrow \min_{\mathbf{u}, z} M \left\{ \left[ \frac{\|\mathbf{y}_{\Omega}\|_2}{\sqrt{M}} z^H \right] \mathbf{T}^{-1}(\mathbf{u}) \left[ \frac{\|\mathbf{y}_{\Omega}\|_2}{\sqrt{M}} z \right] + u_1 \right\}, \text{ subject to } \mathbf{T}(\mathbf{u}) \geq \mathbf{0}, z_{\Omega} = \mathbf{y}_{\Omega} \\
 &\Leftrightarrow \min_{x, \mathbf{u}, z} M \{x + u_1\}, \text{ subject to } \begin{bmatrix} x & \frac{\|\mathbf{y}_{\Omega}\|_2}{\sqrt{M}} z^H \\ \frac{\|\mathbf{y}_{\Omega}\|_2}{\sqrt{M}} z & \mathbf{T}(\mathbf{u}) \end{bmatrix} \geq \mathbf{0}, z_{\Omega} = \mathbf{y}_{\Omega} \\
 &\Leftrightarrow \min_z 2M \left\| \frac{\|\mathbf{y}_{\Omega}\|_2}{\sqrt{M}} z \right\|_{\mathcal{A}}, \text{ subject to } z_{\Omega} = \mathbf{y}_{\Omega} \\
 &\Leftrightarrow \min_z 2\sqrt{M} \|\mathbf{y}_{\Omega}\|_2 \cdot \|z\|_{\mathcal{A}}, \text{ subject to } z_{\Omega} = \mathbf{y}_{\Omega} \\
 &\Leftrightarrow \min_z \|z\|_{\mathcal{A}}, \text{ subject to } z_{\Omega} = \mathbf{y}_{\Omega}.
 \end{aligned} \tag{11.159}$$

This means, like in the ULA case, that GLS is equivalent to ANM subject to the data consistency as if there were no noise.

Finally, note that GLS is practically attractive since it does not require knowledge of the noise level. Regarding this fact, note that GLS is different from ANM in the frequency retrieval process: whereas [Theorem 11.5](#) is used in ANM, [Corollary 11.1](#) is employed in GLS since the noise variance is also encoded in the Toeplitz covariance matrix  $\mathbf{T}(\mathbf{u})$ , besides the frequencies.

### 11.6.4 THE MULTIPLE SNAPSHOT CASE: COVARIANCE FITTING METHODS

In this and the following subsections we will study gridless DOA estimation methods for multiple snapshots. The methods that we will introduce are mainly based on or inspired by those in the single snapshot case in the preceding subsection. The key techniques of these methods exploit the temporal redundancy of the multiple snapshots for possibly improved performance. We have decided to introduce the covariance fitting methods first since they appeared earlier than their deterministic peers. In this kind of methods, differently from the deterministic ones, certain statistical assumptions on the sources (like for SPICE) are required to explicitly express the data covariance matrix. We will discuss three covariance-based gridless sparse methods: GLS in [92], the SMV-based atomic norm method in [149] and the low rank matrix denoising approach in [150]. While GLS is applicable to an arbitrary number of snapshots, the latter two can only be used if there are sufficiently many snapshots.

#### 11.6.4.1 Gridless SPICE (GLS)

In the presence of multiple snapshots, GLS is derived in a similar way as in the single snapshot case by utilizing the convex re-parameterization of the data covariance matrix  $\mathbf{R}$  in Eq. (11.149). For convenience, some derivations of SPICE provided in Section 11.4.5 will be repeated here. We first consider the ULA case. Let  $\tilde{\mathbf{R}} = \frac{1}{L} \mathbf{Y} \mathbf{Y}^H$  be the sample covariance. In the case of  $L \geq M$  when the sample covariance  $\tilde{\mathbf{R}}$  is nonsingular, the following SPICE covariance fitting criterion is minimized:

$$\left\| \mathbf{R}^{-\frac{1}{2}} (\tilde{\mathbf{R}} - \mathbf{R}) \tilde{\mathbf{R}}^{-\frac{1}{2}} \right\|_{\text{F}}^2 = \text{Tr}(\mathbf{R}^{-1} \tilde{\mathbf{R}}) + \text{Tr}(\tilde{\mathbf{R}}^{-1} \mathbf{R}) - 2M. \quad (11.160)$$

Inserting Eq. (11.149) into Eq. (11.160), we have the following equivalences:

$$\begin{aligned} & \min_{\mathbf{u}} \text{Tr} \left( \tilde{\mathbf{R}}^{\frac{1}{2}} \mathbf{T}^{-1}(\mathbf{u}) \tilde{\mathbf{R}}^{\frac{1}{2}} \right) + \text{Tr} \left( \tilde{\mathbf{R}}^{-1} \mathbf{T}(\mathbf{u}) \right), \text{ subject to } \mathbf{T}(\mathbf{u}) \geq \mathbf{0} \\ & \Leftrightarrow \min_{X, \mathbf{u}} \text{Tr}(X) + \text{Tr} \left( \tilde{\mathbf{R}}^{-1} \mathbf{T}(\mathbf{u}) \right), \text{ subject to } \mathbf{T}(\mathbf{u}) \geq \mathbf{0} \text{ and } X \geq \tilde{\mathbf{R}}^{\frac{1}{2}} \mathbf{T}^{-1}(\mathbf{u}) \tilde{\mathbf{R}}^{\frac{1}{2}} \\ & \Leftrightarrow \min_{X, \mathbf{u}} \text{Tr}(X) + \text{Tr} \left( \tilde{\mathbf{R}}^{-1} \mathbf{T}(\mathbf{u}) \right), \text{ subject to } \begin{bmatrix} X & \tilde{\mathbf{R}}^{\frac{1}{2}} \\ \tilde{\mathbf{R}}^{\frac{1}{2}} & \mathbf{T}(\mathbf{u}) \end{bmatrix} \geq \mathbf{0}. \end{aligned} \quad (11.161)$$

Therefore, the covariance fitting problem is cast as SDP. Once the problem is solved, the estimates of the parameters  $(\hat{\mathbf{f}}, \hat{\mathbf{p}}, \hat{\sigma})$  can be obtained from the data covariance estimate  $\hat{\mathbf{R}} = \mathbf{T}(\hat{\mathbf{u}})$  by applying Corollary 11.1.

In the case of  $L < M$  when  $\tilde{\mathbf{R}}$  is singular, we instead minimize the criterion

$$\left\| \mathbf{R}^{-\frac{1}{2}} (\tilde{\mathbf{R}} - \mathbf{R}) \tilde{\mathbf{R}}^{-\frac{1}{2}} \right\|_{\text{F}}^2 = \text{Tr}(\mathbf{R}^{-1} \tilde{\mathbf{R}}^2) + \text{Tr}(\mathbf{R}) - 2\text{Tr}(\tilde{\mathbf{R}}). \quad (11.162)$$

Similarly, inserting Eq. (11.149) into Eq. (11.162), we obtain the following SDP:

$$\min_{X, \mathbf{u}} \text{Tr}(X) + \text{Tr}(\mathbf{T}(\mathbf{u})), \text{ subject to } \begin{bmatrix} X & \tilde{\mathbf{R}} \\ \tilde{\mathbf{R}} & \mathbf{T}(\mathbf{u}) \end{bmatrix} \geq \mathbf{0}. \quad (11.163)$$

The parameter estimates  $(\hat{\mathbf{f}}, \hat{\mathbf{p}}, \hat{\sigma})$  can be obtained in the same manner as above.

The dimensionality of the SDP in Eq. (11.163) can be reduced [84]. To do so, let  $\tilde{\mathbf{Y}} = \frac{1}{L} \mathbf{Y} (\mathbf{Y}^H \mathbf{Y})^{\frac{1}{2}} \in \mathbb{C}^{M \times L}$  and observe that

$$\tilde{\mathbf{R}}^2 = \frac{1}{L^2} \mathbf{Y} \mathbf{Y}^H \mathbf{Y} \mathbf{Y}^H = \tilde{\mathbf{Y}} \tilde{\mathbf{Y}}^H. \quad (11.164)$$

Inserting Eq. (11.164) into Eq. (11.162), we obtain another SDP formulation of the covariance fitting problem:

$$\min_{X, u} \text{Tr}(X) + \text{Tr}(\mathbf{T}(u)), \text{ subject to } \begin{bmatrix} X & \tilde{\mathbf{Y}}^H \\ \tilde{\mathbf{Y}} & \mathbf{T}(u) \end{bmatrix} \geq 0. \quad (11.165)$$

Compared to Eq. (11.163), the dimensionality of the semidefinite matrix in Eq. (11.165) is reduced from  $2M \times 2M$  to  $(M+L) \times (M+L)$  (note that in this case  $L < M$ ). This reduction can be significant in the case of a small number of snapshots.

Similar results can be obtained in the SLA case. In particular, let  $\mathbf{R}_\Omega$  and  $\tilde{\mathbf{R}}_\Omega = \frac{1}{L} \mathbf{Y}_\Omega \mathbf{Y}_\Omega^H$  denote the data covariance and the sample covariance, respectively. Using the linear re-parameterization of  $\mathbf{R}_\Omega$  in Eq. (11.155), similar SDPs to those above can be formulated. In the case of  $L \geq M$  when  $\tilde{\mathbf{R}}_\Omega$  is nonsingular, we have the following SDP:

$$\min_{X, u} \text{Tr}(X) + \text{Tr}\left(\boldsymbol{\Gamma}_\Omega^T \tilde{\mathbf{R}}_\Omega^{-1} \boldsymbol{\Gamma}_\Omega \mathbf{T}(u)\right), \text{ subject to } \begin{bmatrix} X & \tilde{\mathbf{R}}_\Omega^{\frac{1}{2}} \\ \tilde{\mathbf{R}}_\Omega^{\frac{1}{2}} & \boldsymbol{\Gamma}_\Omega \mathbf{T}(u) \boldsymbol{\Gamma}_\Omega^T \end{bmatrix} \geq 0. \quad (11.166)$$

When  $L < M$ , the SDP is

$$\min_{X, u} \text{Tr}(X) + \text{Tr}(\boldsymbol{\Gamma}_\Omega^T \boldsymbol{\Gamma}_\Omega \mathbf{T}(u)), \text{ subject to } \begin{bmatrix} X & \tilde{\mathbf{R}}_\Omega \\ \tilde{\mathbf{R}}_\Omega & \boldsymbol{\Gamma}_\Omega \mathbf{T}(u) \boldsymbol{\Gamma}_\Omega^T \end{bmatrix} \geq 0, \quad (11.167)$$

where  $\tilde{\mathbf{R}}_\Omega$  can also be replaced by  $\frac{1}{L} \mathbf{Y}_\Omega (\mathbf{Y}_\Omega^H \mathbf{Y}_\Omega)^{\frac{1}{2}} \in \mathbb{C}^{M \times L}$  for dimensionality reduction. Once the SDP is solved, the parameters of interest can be retrieved as in the ULA case.

As in the single snapshot case, GLS is guaranteed to produce a sparse solution with at most  $N - 1$  sources. Besides this, GLS has other attractive properties as detailed below.

Let us assume that the antenna array is a redundancy array or, equivalently, that the full matrix  $\mathbf{T}(u)$  can be recovered from its submatrix  $\boldsymbol{\Gamma}_\Omega \mathbf{T}(u) \boldsymbol{\Gamma}_\Omega^T$  (see, e.g., [151, 152]). Note that all ULAs and many SLAs are redundancy arrays. Then the GLS estimator is statistically consistent in the snapshot number  $L$  if the true source number  $K \leq N - 1$ . To see this, let us consider a ULA first. As  $L \rightarrow \infty$ ,  $\tilde{\mathbf{R}}$  approaches the true data covariance matrix that is denoted by  $\mathbf{R}^o$  and has the same Toeplitz structure as  $\mathbf{R}$ . Hence, it follows from Eq. (11.160) that  $\hat{\mathbf{R}} = \mathbf{R}^o$  and the true parameters can be retrieved from  $\hat{\mathbf{R}}$ . In the SLA case, similarly, the covariance matrix estimate  $\hat{\mathbf{R}}_\Omega$  converges to the true one

as  $L \rightarrow \infty$ . Then, the assumption of redundancy array can be used to show that all the information in the Toeplitz matrix  $\mathbf{T}(\mathbf{u})$  for frequency retrieval can be recovered from  $\hat{\mathbf{R}}_\Omega$  and hence that the true parameters can be obtained.

Furthermore, under the stronger assumption of Gaussian sources and noise, GLS is an asymptotic ML estimator when  $K \leq N - 1$  and a redundancy array is employed. This is true because the global solution of the SPICE covariance fitting problem is a large-snapshot realization of the ML estimator [87, 153] and because GLS globally solves the problem. As a direct consequence of this, GLS has improved resolution as  $L$  increases and the resolution limit vanishes as  $L$  grows to infinity. Another consequence is that GLS can estimate more sources than the number of antennas. In fact, a redundancy array with aperture  $N$  can usually be formed by using  $M \approx \sqrt{3(N - 1)}$  antennas [151]. This means that up to about  $\frac{1}{3}M^2$  sources can be estimated using GLS with only  $M$  antennas.

It is worth noting that generally the above properties of GLS are not shared by its discrete version, viz. SPICE, due to an ambiguity problem, even when the on-grid assumption of SPICE holds. To see this, let us consider the ULA case as an example and suppose that SPICE can accurately estimate the data covariance matrix  $\mathbf{R} = \mathbf{T}(\mathbf{u})$ , as GLS does. Note that when  $\mathbf{R}$  has full rank, which is typically the case in practice, the solution of GLS is provided by the unique decomposition of  $\mathbf{R}$  from Corollary 11.1. But this uniqueness cannot be guaranteed in SPICE according to Remark 11.1 (note that the condition that  $r < N$  of Corollary 11.1 might not be satisfied in SPICE).

#### 11.6.4.2 SMV-based atomic norm minimization (ANM-SMV)

Within the SMV superresolution framework introduced in [100], an ANM approach was proposed in [149], designated here as ANM-SMV. While the paper [149] focused on co-prime arrays [154], which form a special class of SLAs, ANM-SMV can actually be applied to a broader class of SLAs such as redundancy arrays or even general SLAs. To simplify our discussions, we consider without loss of generality a redundancy array, denoted by  $\Omega$ .

Under the assumption of uncorrelated sources, as for GLS, the data covariance matrix  $\mathbf{R}_\Omega$  can be expressed as:

$$\mathbf{R}_\Omega = \mathbf{A}_\Omega(f) \text{diag}(\mathbf{p}) \mathbf{A}_\Omega^H(f) + \sigma \mathbf{I}, \quad (11.168)$$

where  $p_k > 0$ ,  $k = 1, \dots, K$  denote the source powers. According to the discussions in Section 11.6.3.6,  $\mathbf{R}_\Omega$  is a submatrix of a Toeplitz covariance matrix:

$$\mathbf{R}_\Omega = \mathbf{\Gamma}_\Omega \mathbf{T}(\mathbf{u}) \mathbf{\Gamma}_\Omega^T + \sigma \mathbf{I}, \quad (11.169)$$

where  $\mathbf{u} \in \mathbb{C}^N$  and  $\mathbf{T}(\mathbf{u}) = \mathbf{A}(f) \text{diag}(\mathbf{p}) \mathbf{A}^H(f)$ . Since the frequencies have been encoded in  $\mathbf{T}(\mathbf{u})$ , ANM-SMV, like GLS, also carries out covariance fitting to estimate  $\mathbf{u}$  and thus the frequencies. But the technique used by ANM-SMV is different. To describe this technique, note that  $\mathbf{u}$  can be expressed as

$$\mathbf{u} = \mathbf{A}^*(f) \mathbf{p}. \quad (11.170)$$

Let us define  $\mathbf{v} \in \mathbb{C}^{2N-1}$  such that

$$v_j = \begin{cases} u_{N-j+1}, & j = 1, \dots, N, \\ u_{j-N+1}^*, & j = N+1, \dots, 2N-1. \end{cases} \quad (11.171)$$

Given  $\mathbf{u}$  in Eq. (11.170), note that  $\mathbf{v}$  can be viewed as a snapshot of a virtual  $(2N-1)$ -element ULA on which  $K$  sources impinge. Based on this observation, ANM-SMV attempts to solve the following ANM problem:

$$\min_{\tilde{\mathbf{v}}} \|\mathbf{v}\|_{\mathcal{A}}, \text{ subject to } \|\mathbf{v} - \tilde{\mathbf{v}}\|_2 \leq \eta, \quad (11.172)$$

where  $\tilde{\mathbf{v}}$  denotes an estimate of  $\mathbf{v}$ , which will be given later, and  $\eta$  is an upper bound on the error of  $\tilde{\mathbf{v}}$ . By casting Eq. (11.172) as SDP, this problem can be solved and the frequencies can then be estimated as those composing the solution  $\mathbf{v}$ .

The remaining task is to compose the estimate  $\tilde{\mathbf{v}}$  and analyze its error bound  $\eta$ . To do so, the noise variance  $\sigma$  is assumed to be known. Using Eq. (11.169), an estimate of  $\mathbf{\Gamma}_{\Omega} \mathbf{T}(\mathbf{u}) \mathbf{\Gamma}_{\Omega}^T$  is formed as  $\tilde{\mathbf{R}}_{\Omega} - \sigma \mathbf{I}$ . After that, an estimate of  $\mathbf{u}$  is obtained by averaging the corresponding entries of the estimate  $\tilde{\mathbf{R}}_{\Omega} - \sigma \mathbf{I}$ . This can be done since  $\Omega$  is assumed to be a redundancy array. Finally,  $\tilde{\mathbf{v}}$  is obtained by using Eq. (11.171). Under the assumption of i.i.d. Gaussian sources and noise and assuming sufficiently many snapshots, an error bound  $\eta \propto \sigma$  is provided in [149] in the case of co-prime arrays. This bound might be extended to other cases but that is beyond the scope of this article. Based on the above observations and the result in [155], it can be shown that ANM-SMV can stably estimate the frequencies, provided that they are sufficiently separated, with a probability that increases with the snapshot number  $L$ .

#### 11.6.4.3 Nuclear norm minimization followed by MUSIC (NNM-MUSIC)

Using a low rank matrix recovery technique and a subspace method, another covariance-based method was proposed in [150] that is called NNM-MUSIC. Based on Eq. (11.169), the paper [150] proposed a two-step approach: (1) First estimate the full data covariance matrix  $\mathbf{T}(\mathbf{u})$  by exploiting its low rank structure, and (2) Then estimate the frequencies from  $\mathbf{T}(\mathbf{u})$  using MUSIC.

In the first step, the following NNM problem is solved to estimate  $\mathbf{T}(\mathbf{u})$ :

$$\min_{\mathbf{R}} \|\mathbf{R}\|_{\star}, \text{ subject to } \|\mathbf{R} - \mathbf{T}(\tilde{\mathbf{u}})\|_{\text{F}} \leq \eta. \quad (11.173)$$

In Eq. (11.173),  $\mathbf{T}(\tilde{\mathbf{u}})$  denotes an estimate of the data covariance matrix, which is obtained by averaging the corresponding entries of the sample covariance matrix  $\tilde{\mathbf{R}}_{\Omega}$ , and  $\eta$  measures the distance between the true low rank matrix  $\mathbf{T}(\mathbf{u})$  and the above estimate in the Frobenius norm. Once  $\mathbf{R}$  is solved for, MUSIC is adopted to estimate the frequencies from the subspace of  $\mathbf{R}$ .

By setting  $\eta = \sqrt{N}\sigma$  and assuming  $L \rightarrow \infty$ , it was shown in [150] that solving Eq. (11.173) can exactly recover  $\mathbf{T}(\mathbf{u})$ . However, we note that it is not an easy task to choose  $\eta$  in practice. Moreover, although the Toeplitz structure embedded in the

data covariance matrix  $\mathbf{R}_\Omega$  is exploited to form the estimate  $\tilde{\mathbf{u}}$ , this structure is not utilized in the matrix denoising step. It was argued in [150] that the PSDness of  $\mathbf{R}$  can be preserved by solving Eq. (11.173) if  $\mathbf{T}(\tilde{\mathbf{u}})$  is PSD, which holds true if sufficiently many snapshots are available.

#### 11.6.4.4 Comparison of GLS, ANM-SMV, and NNM-MUSIC

In this subsection we compare the three covariance-based methods, namely GLS, ANM-SMV, and NNM-MUSIC. While these methods are derived under similar assumptions on sources and noise, we argue that GLS can outperform ANM-SMV and NNM-MUSIC in several aspects. First, GLS is hyperparameter-free and can consistently estimate the noise variance  $\sigma$ , whereas ANM-SMV and NNM-MUSIC usually require knowledge of this variance since the error bounds  $\eta$  in Eqs. (11.172), (11.173) are functions of  $\sigma$ . In fact, even when  $\sigma$  is known the choice of  $\eta$  is still not easy. Second, ANM-SMV and NNM-MUSIC are usable only with sufficiently many snapshots (which are needed to obtain a reasonable estimate of  $\mathbf{u}$  as well as a reasonable error bound  $\eta$ ), while GLS can be used even with a single snapshot. Third, GLS and NNM-MUSIC are statistically consistent but ANM-SMV is not. Note that ANM-SMV still suffers from a resolution limit even if  $\tilde{\mathbf{v}}$  in Eq. (11.172) is exactly estimated given an infinitely number of snapshots. Fourth, GLS is a large-snapshot realization of the ML estimation while ANM-SMV and NNM-MUSIC are not.

Last but not least, ANM-SMV and NNM-MUSIC cannot exactly recover the frequencies in the absence of noise since the estimate  $\tilde{\mathbf{v}}$  or  $\tilde{\mathbf{u}}$  will suffer from some approximation error with finite snapshots. In contrast to this, GLS can exactly recover the frequencies under a mild frequency separation condition due to its connection to the atomic norm that will be discussed in Section 11.6.5.3 (considering the single snapshot case as an example).

### 11.6.5 THE MULTIPLE SNAPSHOT CASE: DETERMINISTIC METHODS

In this subsection we present several deterministic gridless sparse methods for the case of multiple snapshots. The main idea is to utilize the temporal redundancy of the snapshots. Different from the covariance-based methods in the preceding subsection, these deterministic optimization methods are derived without statistical assumptions on the sources (though weak technical assumptions might be needed to provide theoretical guarantees). As in the single snapshot case, we first provide a general framework for such methods. We then discuss the potential advantages of multiple snapshots for DOA/frequency estimation based on an MMV atomic  $\ell_0$  norm formulation. After that, the MMV atomic norm method will be presented. Finally, a possible extension of EMaC to the multiple snapshot case is discussed.

#### 11.6.5.1 A general framework

The data model in Eq. (11.99) can be written as:

$$\mathbf{Y}_\Omega = \mathbf{Z}_\Omega + \mathbf{E}_\Omega, \quad \mathbf{Z} = \mathbf{A}(f)\mathbf{S}, \quad (11.174)$$

where  $\mathbf{Z}$  denotes the noiseless multiple snapshot signal that contains the frequencies of interest. In the presence of bounded noise with  $\|\mathbf{E}_\Omega\|_F \leq \eta$ , we solve the constrained optimization problem:

$$\min_{\mathbf{Z}} \mathcal{M}(\mathbf{Z}), \text{ subject to } \|\mathbf{Z}_\Omega - \mathbf{Y}_\Omega\|_F \leq \eta. \quad (11.175)$$

We can instead solve the regularized optimization problem given by

$$\min_{\mathbf{Z}} \lambda \mathcal{M}(\mathbf{Z}) + \frac{1}{2} \|\mathbf{Z}_\Omega - \mathbf{Y}_\Omega\|_F^2, \quad (11.176)$$

where  $\lambda > 0$  is a regularization parameter. In the extreme noiseless case, both Eqs. (11.175) and (11.176) degenerate to the following problem:

$$\min_{\mathbf{Z}} \mathcal{M}(\mathbf{Z}), \text{ subject to } \mathbf{Z}_\Omega = \mathbf{Y}_\Omega. \quad (11.177)$$

In Eqs. (11.175)–(11.177),  $\mathcal{M}(\mathbf{Z})$  denotes a sparse metric. By minimizing  $\mathcal{M}(\mathbf{Z})$  we attempt to reduce the number of frequencies composing  $\mathbf{Z}$ .

### 11.6.5.2 Atomic $\ell_0$ norm

In this subsection we study the advantage of using multiple snapshots for frequency estimation and extend the result in Section 11.4.3.1 from the discrete to the continuous setting. To do so, we extend the atomic  $\ell_0$  norm from the single to the multiple snapshot case. Note that the noiseless multiple snapshot signal  $\mathbf{Z}$  in Eq. (11.174) can be written as:

$$\mathbf{Z} = \sum_{k=1}^K \mathbf{a}(f_k) s_k = \sum_{k=1}^K c_k \mathbf{a}(f_k) \boldsymbol{\phi}_k, \quad (11.178)$$

where  $s_k = [s_{k1}, \dots, s_{kL}] \in \mathbb{C}^{1 \times L}$ ,  $c_k = \|s_k\|_2 > 0$ , and  $\boldsymbol{\phi}_k = c_k^{-1} s_k \in \mathbb{C}^{1 \times L}$  with  $\|\boldsymbol{\phi}_k\|_2 = 1$ . We therefore define the set of atoms in this case as:

$$\mathcal{A} = \{\mathbf{a}(f_k, \boldsymbol{\phi}_k) = \mathbf{a}(f_k) \boldsymbol{\phi}_k : f_k \in \mathbb{T}, \boldsymbol{\phi}_k \in \mathbb{C}^{1 \times L}, \|\boldsymbol{\phi}_k\|_2 = 1\} \quad (11.179)$$

Note that  $\mathbf{Z}$  is a linear combination of  $K$  atoms in  $\mathcal{A}$ . The atomic  $\ell_0$  norm of  $\mathbf{Z}$  induced by the new atomic set  $\mathcal{A}$  is given by:

$$\begin{aligned} \|\mathbf{Z}\|_{\mathcal{A},0} &= \inf_{c_k, f_k, \boldsymbol{\phi}_k} \left\{ \mathcal{K} : \mathbf{Z} = \sum_{k=1}^K \mathbf{a}(f_k, \boldsymbol{\phi}_k) c_k, f_k \in \mathbb{T}, \|\boldsymbol{\phi}_k\|_2 = 1, c_k > 0 \right\} \\ &= \inf_{f_k, s_k} \left\{ \mathcal{K} : \mathbf{Z} = \sum_{k=1}^K \mathbf{a}(f_k) s_k, f_k \in \mathbb{T} \right\}. \end{aligned} \quad (11.180)$$

Using the atomic  $\ell_0$  norm, in the noiseless case, the problem resulting from Eq. (11.177) is given by:

$$\min_{\mathbf{Z}} \|\mathbf{Z}\|_{\mathcal{A},0}, \text{ subject to } \mathbf{Z}_\Omega = \mathbf{Y}_\Omega. \quad (11.181)$$

To show the advantage of multiple snapshots, we define the continuous dictionary (see also Eq. 11.7)

$$\mathcal{A}_\Omega^1 = \{\mathbf{a}_\Omega(f) : f \in \mathbb{T}\} \quad (11.182)$$

and let  $\text{spark}(\mathcal{A}_\Omega^1)$  be its spark. We have the following theoretical guarantee for Eq. (11.181) that generalizes the result in [20, Theorem 2.4].

**Theorem 11.12 ([22])**  $\mathbf{Y} = \sum_{j=1}^K c_j \mathbf{a}(f_j, \boldsymbol{\phi}_j)$  is the unique solution to Eq. (11.181) if

$$K < \frac{\text{spark}(\mathcal{A}_\Omega^1) - 1 + \text{rank}(\mathbf{Y}_\Omega)}{2}. \quad (11.183)$$

Moreover, the atomic decomposition above is the only one that satisfies  $\|\mathbf{Y}\|_{\mathcal{A},0} = K$ .

Note that the condition in Eq. (11.183) coincides with that in Eq. (11.10) required to guarantee parameter identifiability for DOA estimation. In fact it can be shown that, in the noiseless case, the frequencies/DOAs can be uniquely determined by the atomic  $\ell_0$  norm optimization if and only if they can be uniquely identified (see, e.g., [156, Proposition 2]). Furthermore, the above result also holds true for general array geometries and even for general parameter estimation problems, provided that the atomic  $\ell_0$  norm is appropriately defined.

By Theorem 11.12 the frequencies can be exactly determined by Eq. (11.181) if the number of sources  $K$  is sufficiently small with respect to the array geometry  $\Omega$  and the observed data  $\mathbf{Y}_\Omega$ . Note that the number of recoverable frequencies can be increased, as compared to the SMV case, if  $\text{rank}(\mathbf{Y}_\Omega) > 1$ , which always happens but in the trivial case when the multiple snapshots in  $\mathbf{Y}_\Omega$  are identical up to scaling factors.

As in the single snapshot case, computing  $\|\mathbf{Z}\|_{\mathcal{A},0}$  can be cast as a rank minimization problem. To see this, let  $\mathbf{T}(\mathbf{u})$  be a Toeplitz matrix and impose the condition that

$$\begin{bmatrix} \mathbf{X} & \mathbf{Z}^H \\ \mathbf{Z} & \mathbf{T}(\mathbf{u}) \end{bmatrix} \geq 0, \quad (11.184)$$

where  $\mathbf{X}$  is a free matrix variable. By invoking the Vandermonde decomposition, we see that  $\mathbf{T}(\mathbf{u})$  admits a  $\text{rank}(\mathbf{T}(\mathbf{u}))$ -atomic Vandermonde decomposition. Moreover,  $\mathbf{Z}$  lies in the range space of  $\mathbf{T}(\mathbf{u})$  and thus it can be expressed by  $\text{rank}(\mathbf{T}(\mathbf{u}))$  atoms. Formally, we have the following result.

**Theorem 11.13 ([22, 157])**  $\|\mathbf{Z}\|_{\mathcal{A},0}$  defined in Eq. (11.180) equals the optimal value of the following rank minimization problem:

$$\min_{\mathbf{X}, \mathbf{u}} \text{rank}(\mathbf{T}(\mathbf{u})), \text{ subject to Eq. (11.184)}. \quad (11.185)$$

While the rank minimization problem cannot be easily solved, we next discuss its convex relaxation, namely the atomic norm method.

### 11.6.5.3 Atomic norm

To provide a practical approach, we study the atomic norm induced by the atomic set  $\mathcal{A}$  defined in Eq. (11.179). As in the single snapshot case, we have that

$$\begin{aligned}
\|\mathbf{Z}\|_{\mathcal{A}} &= \inf\{t > 0 : \mathbf{Z} \in t\text{conv}(\mathcal{A})\} \\
&= \inf_{c_k, f_k, \boldsymbol{\phi}_k} \left\{ \sum_k c_k : \mathbf{Z} = \sum_k \mathbf{a}(f_k, \boldsymbol{\phi}_k) c_k, f_k \in \mathbb{T}, \|\boldsymbol{\phi}_k\|_2 = 1, c_k > 0 \right\} \\
&= \inf_{f_k, \mathbf{s}_k} \left\{ \sum_k \|\mathbf{s}_k\|_2 : \mathbf{Z} = \sum_k \mathbf{a}(f_k) \mathbf{s}_k, f_k \in \mathbb{T} \right\}.
\end{aligned} \tag{11.186}$$

Note that  $\|\mathbf{Z}\|_{\mathcal{A}}$  is in fact a continuous counterpart of the  $\ell_{2,1}$  norm. Moreover, it is shown in the following result that  $\|\mathbf{Z}\|_{\mathcal{A}}$  can also be cast as SDP.

**Theorem 11.14** ([22, 85])  $\|\mathbf{Z}\|_{\mathcal{A}}$  defined in Eq. (11.186) equals the optimal value of the following SDP:

$$\min_{X, \mathbf{u}} \frac{1}{2\sqrt{N}} [\text{Tr}(X) + \text{Tr}(\mathbf{T}(\mathbf{u}))], \text{ subject to Eq. (11.184).} \tag{11.187}$$

The proof of Theorem 11.14 is omitted since it is very similar to that in the case of a single snapshot. Similarly, the frequencies and the powers are encoded in the Toeplitz matrix  $\mathbf{T}(\mathbf{u})$  and therefore, once  $\mathbf{T}(\mathbf{u})$  is obtained, they can be retrieved from its Vandermonde decomposition.

By applying Theorem 11.14, in the noiseless case, the ANM problem resulting from Eq. (11.177) can be cast as the following SDP:

$$\min_{X, \mathbf{u}, \mathbf{Z}} \text{Tr}(X) + \text{Tr}(\mathbf{T}(\mathbf{u})), \text{ subject to Eq. (11.184) and } \mathbf{Z}_{\Omega} = \mathbf{Y}_{\Omega}. \tag{11.188}$$

For Eq. (11.188), we have theoretical guarantees similar to those in the single snapshot case; in other words, the frequencies can be exactly recovered by solving Eq. (11.188) under appropriate conditions. Formally, we have the following results that generalize those in the single snapshot case.

**Theorem 11.15** ([22])  $\mathbf{Y} = \sum_{j=1}^K c_j \mathbf{a}(f_j, \boldsymbol{\phi}_j)$  is the unique atomic decomposition satisfying  $\|\mathbf{Y}\|_{\mathcal{A}} = \sum_{j=1}^K c_j$  if  $\Delta_T \geq \frac{1}{\lfloor (N-1)/4 \rfloor}$  and  $N \geq 257$ .

**Theorem 11.16** ([22]) Suppose we observe the rows of  $\mathbf{Y} = \sum_{j=1}^K c_j \mathbf{a}(f_j, \boldsymbol{\phi}_j)$  that are indexed by  $\Omega$ , where  $\Omega \subset [1, \dots, N]$  is of size  $M$  and is selected uniformly at random. Assume that  $\{\boldsymbol{\phi}_j\}$  are independent random variables on the unit hypersphere with  $\mathbb{E} \boldsymbol{\phi}_j = \mathbf{0}$ . If  $\Delta_T \geq \frac{1}{\lfloor (N-1)/4 \rfloor}$ , then there exists a numerical constant  $C$  such that

$$M \geq C \max \left\{ \log^2 \frac{\sqrt{LN}}{\delta}, K \log \frac{K}{\delta} \log \frac{\sqrt{LN}}{\delta} \right\} \tag{11.189}$$

is sufficient to guarantee that, with probability at least  $1 - \delta$ ,  $\mathbf{Y}$  is the unique solution to Eq. (11.188) and  $\mathbf{Y} = \sum_{j=1}^K c_j \mathbf{a}(f_j, \boldsymbol{\phi}_j)$  is the unique atomic decomposition satisfying  $\|\mathbf{Y}\|_{\mathcal{A}} = \sum_{j=1}^K c_j$ .

Note that we have not made any assumptions on the sources in [Theorem 11.15](#) and therefore it applies to all kinds of source signals, including coherent sources. As a result, one cannot expect that the theoretical guarantee improves over the single snapshot case.

Note that the dependence of  $M$  on  $L$  in the bound [\(11.189\)](#) is for controlling the probability of successful recovery. To make it clear, consider the case when we seek to recover the columns of  $\mathbf{Y}$  independently using the single snapshot ANM method. When  $M$  satisfies Eq. [\(11.189\)](#) for  $L = 1$ , each column of  $\mathbf{Y}$  can be recovered with probability  $1 - \delta$ . It follows that  $\mathbf{Y}$  can be exactly recovered with probability at least  $1 - L\delta$ . In contrast to this, if we recover  $\mathbf{Y}$  using the multiple snapshot ANM method, then with the same number of measurements the success probability is improved to  $1 - \sqrt{L}\delta$  (to see this, replace  $\delta$  in Eq. [\(11.189\)](#) by  $\sqrt{L}\delta$ ).

Note also that the assumption on  $\{\boldsymbol{\phi}_j\}$  in [Theorem 11.16](#) is weak in the sense that the sources can be coherent. To see this, suppose that the rows of  $\mathbf{S}$  are i.i.d. Gaussian with mean zero and a covariance matrix whose rank is equal to one. Then the sources are identical up to random global phases and hence they are independent but coherent. This explains why the theoretical guarantee given in [Theorem 11.16](#) does not improve over the similar result in the single snapshot case. In other words, the results of Theorems 11.15 and 11.16 are *worst case* analysis. Although these results do not shed light on the advantage of multiple snapshots, numerical simulations show that the atomic norm approach significantly improves the recovery performance, compared to the case of  $L = 1$ , when the source signals are at general positions (see, e.g., [\[22, 85, 157\]](#)).

In the presence of noise, the ANM problem resulting from Eq. [\(11.176\)](#) can be formulated as the following SDP:

$$\min_{X, u, Z} \frac{\lambda}{2\sqrt{N}} [\text{Tr}(X) + \text{Tr}(\mathbf{T}(u))] + \frac{1}{2} \|Z_\Omega - Y_\Omega\|_F^2, \text{ subject to Eq. } (11.184). \quad (11.190)$$

It was shown in [\[85\]](#) that in the ULA case the choice of  $\lambda \approx \sqrt{M(L + \log M + \sqrt{2L \log M})}\sigma$  results in a stable recovery of the signal  $Z$ , which generalizes the result in the single snapshot case. In the SLA case, a similar parameter tuning method can be derived by combining the techniques in [\[85, 133\]](#).

#### 11.6.5.4 Hankel-based nuclear norm

In this subsection we extend the EMaC method from the single to the multiple snapshot case. For each noiseless snapshot  $z(t)$ , we can form an  $m \times n$  Hankel matrix  $\mathbf{H}(z(t))$  as in Eq. [\(11.138\)](#), where  $m + n = N + 1$ , and then combine them in the following  $m \times nL$  matrix:

$$\mathbf{H}(Z) = [\mathbf{H}(z(1)), \dots, \mathbf{H}(z(L))]. \quad (11.191)$$

Using the decomposition in Eq. [\(11.139\)](#), we have that

$$\begin{aligned} \mathbf{H}(\mathbf{Z}) &= \sum_{k=1}^K \begin{bmatrix} 1 \\ e^{i2\pi f_k} \\ \vdots \\ e^{i2\pi(m-1)f_k} \end{bmatrix} \\ &\times \begin{bmatrix} s_k(1), s_k(1)e^{i2\pi f_k}, \dots, s_k(1)e^{i2\pi(n-1)f_k}, \dots, s_k(L), s_k(L)e^{i2\pi f_k}, \dots, s_k(L)e^{i2\pi(n-1)f_k} \end{bmatrix}. \end{aligned} \quad (11.192)$$

As a result,

$$\text{rank}(\mathbf{H}(\mathbf{Z})) \leq K. \quad (11.193)$$

It follows that if  $K < \min(m, nL)$ , then  $\mathbf{H}(\mathbf{Z})$  is a low rank matrix. Therefore, we may recover  $\mathbf{H}(\mathbf{Z})$  by minimizing its nuclear norm, i.e., by letting (see Eq. 11.142)

$$\begin{aligned} \mathcal{M}(\mathbf{Z}) &= \|\mathbf{H}(\mathbf{Z})\|_* \\ &= \min_{\mathbf{Q}_1, \mathbf{Q}_2} \frac{1}{2} [\text{Tr}(\mathbf{Q}_1) + \text{Tr}(\mathbf{Q}_2)], \text{ subject to } \begin{bmatrix} \mathbf{Q}_1 & \mathbf{H}(\mathbf{Z})^H \\ \mathbf{H}(\mathbf{Z}) & \mathbf{Q}_2 \end{bmatrix} \geq \mathbf{0}. \end{aligned} \quad (11.194)$$

The resulting approach is referred to as M-EMaC.

A challenging problem when applying M-EMaC is the choice of the parameter  $m$ . Intuitively, we need to ensure that the equality holds in Eq. (11.193) for the true signal  $\mathbf{Z}$  so that the data information can be appropriately encoded and the frequencies can be correctly recovered from  $\mathbf{H}(\mathbf{Z})$ . This is guaranteed for a single snapshot once  $\mathbf{H}(\mathbf{Z})$  is rank deficient. Unfortunately, a similar argument does not hold in the case of multiple snapshots. In particular, it can be seen from Eq. (11.192) that the rank of  $\mathbf{H}(\mathbf{Z})$  also depends on the unknown source signals  $\mathbf{S}$ . As an example, in the extreme case when all the sources are coherent, we have that

$$\text{rank}(\mathbf{H}(\mathbf{Z})) \leq \min(K, m, n), \quad (11.195)$$

which can be much smaller than  $nL$ . As a result, if we know that  $K < \frac{N}{2}$ , then we can set  $m = \left\lceil \frac{N}{2} \right\rceil$ . We leave the parameter tuning in the case of  $K \geq \frac{N}{2}$  as an open problem.

### 11.6.6 REWEIGHTED ATOMIC NORM MINIMIZATION

In contrast to the EMaC, the atomic norm methods of Sections 11.6.3.3 and 11.6.5.3 can better preserve the signal structure, which is important especially in the noisy case. However, a major disadvantage of the atomic norm is that it can theoretically guarantee successful frequency recovery if they are separated by at least  $\frac{2.52}{N}$  [136]. To overcome this “resolution limit,”<sup>2</sup> we present in this subsection the reweighted atomic-norm minimization (RAM) method that was proposed in [86]. The key idea

---

<sup>2</sup>Since the aforementioned frequency separation condition is sufficient but not necessary, this might not be the real resolution limit of the atomic norm.

of RAM is to use a smooth surrogate for the atomic  $\ell_0$  norm, which exploits the sparsity to the greatest extent possible and does not suffer from any resolution limit but is nonconvex and nonsmooth, and then optimize the surrogate using a reweighted approach. Interestingly, the resulting reweighted approach is shown to be a reweighted atomic norm with a sound weighting function that gradually enhances sparsity and resolution. While several reweighted approaches have been proposed in the discrete setting (see, e.g., [60, 158–160]), RAM appears to be the first for continuous dictionaries. Since RAM can be applied to single or multiple snapshots as well as to ULA or SLA, we present the result in the most general multiple snapshot SLA case, as in the preceding subsection.

#### 11.6.6.1 A smooth surrogate for $\|\mathbf{Z}\|_{\mathcal{A},0}$

To derive RAM, we first introduce a smooth surrogate for  $\|\mathbf{Z}\|_{\mathcal{A},0}$  defined in Eq. (11.180). Note that if the surrogate function is given directly in the continuous frequency domain, then a difficult question is whether and how it can be formulated as a finite-dimensional optimization problem, as  $\|\mathbf{Z}\|_{\mathcal{A},0}$  and  $\|\mathbf{Z}\|_{\mathcal{A}}$ . To circumvent this problem, RAM operates instead in the re-parameterized  $\mathbf{u}$  domain. In particular, we have shown that  $\|\mathbf{Z}\|_{\mathcal{A},0}$  is equivalent to the following rank minimization problem:

$$\min_{X, \mathbf{u}} \text{rank}(\mathbf{T}(\mathbf{u})), \text{ subject to } \begin{bmatrix} X & \mathbf{Z}^H \\ \mathbf{Z} & \mathbf{T}(\mathbf{u}) \end{bmatrix} \geq \mathbf{0}. \quad (11.196)$$

Inspired by the literature on low rank matrix recovery (see, e.g., [161–163]), the log-det heuristic is adopted as a smooth surrogate for the matrix rank, resulting in the following sparse metric:

$$\mathcal{M}^\epsilon(\mathbf{Z}) = \min_{X, \mathbf{u}} \log |\mathbf{T}(\mathbf{u}) + \epsilon I| + \text{Tr}(X), \text{ subject to } \begin{bmatrix} X & \mathbf{Z}^H \\ \mathbf{Z} & \mathbf{T}(\mathbf{u}) \end{bmatrix} \geq \mathbf{0}. \quad (11.197)$$

In Eq. (11.197),  $\epsilon$  is a tuning parameter that avoids the first term in the objective from being  $-\infty$  when  $\mathbf{T}(\mathbf{u})$  is rank-deficient, and  $\text{Tr}(X)$  is included in the objective to prevent the trivial solution  $\mathbf{u} = \mathbf{0}$ , as in the SDP in Eq. (11.187) for the atomic norm. Intuitively, as  $\epsilon \rightarrow 0$ , the log-det heuristic approaches the rank function and  $\mathcal{M}^\epsilon(\mathbf{Z})$  would approach  $\|\mathbf{Z}\|_{\mathcal{A},0}$ . This is indeed true and is formally stated in the following result.

**Theorem 11.17** ([86]) *Let  $r = \|\mathbf{Z}\|_{\mathcal{A},0}$  and let  $\epsilon \rightarrow 0$ . Then, we have the following results:*

1. *If  $r \leq N - 1$ , then*

$$\mathcal{M}^\epsilon(\mathbf{Z}) \sim (r - N) \ln \frac{1}{\epsilon}, \quad (11.198)$$

*i.e., the two quantities above are equivalent infinities. Otherwise,  $\mathcal{M}^\epsilon(\mathbf{Z})$  approaches a constant depending only on  $\mathbf{Z}$ ;*

2. *Let  $\hat{\mathbf{u}}_\epsilon$  be the (global) solution  $\mathbf{u}$  to the optimization problem in Eq. (11.197). Then, the smallest  $N - r$  eigenvalues of  $\mathbf{T}(\hat{\mathbf{u}}_\epsilon)$  are either zero or approach zero as fast as  $\epsilon$ ;*

3. For any cluster point of  $\hat{\mathbf{u}}_\epsilon$  at  $\epsilon = 0$ , denoted by  $\hat{\mathbf{u}}_0$ , there exists an atomic decomposition  $\mathbf{Z} = \sum_{k=1}^r \mathbf{a}(f_k) \mathbf{s}_k$  such that  $\mathbf{T}(\hat{\mathbf{u}}_0) = \sum_{k=1}^r \|\mathbf{s}_k\|_2^2 \mathbf{a}(f_k) \mathbf{a}(f_k)^H$ .<sup>3</sup>

Theorem 11.17 shows that the sparse metric  $\mathcal{M}^\epsilon(\mathbf{Z})$  approaches  $\|\mathbf{Z}\|_{\mathcal{A},0}$  as  $\epsilon \rightarrow 0$ . Moreover, it characterizes the properties of the optimizer  $\hat{\mathbf{u}}_\epsilon$ , as  $\epsilon \rightarrow 0$ , including the convergence speed of the smallest  $N - \|\mathbf{Z}\|_{\mathcal{A},0}$  eigenvalues and the limiting form of  $\mathbf{T}(\hat{\mathbf{u}}_0)$  via the Vandermonde decomposition. It is worth noting that the term  $\ln \frac{1}{\epsilon}$  in Eq. (11.198), which becomes unbounded as  $\epsilon \rightarrow 0$ , is not problematic in the optimization problem, since the objective function  $\mathcal{M}^\epsilon(\mathbf{Z})$  can be re-scaled by  $\left( \ln \frac{1}{\epsilon} \right)^{-1}$  for any  $\epsilon > 0$  without altering the optimizer.

In another interesting extreme case when  $\epsilon \rightarrow +\infty$ , the following result shows that  $\mathcal{M}^\epsilon(\mathbf{Z})$  in fact plays the same role as  $\|\mathbf{Z}\|_{\mathcal{A}}$ .

**Theorem 11.18** ([86]) Let  $\epsilon \rightarrow +\infty$ . Then,

$$\mathcal{M}^\epsilon(\mathbf{Z}) - N \ln \epsilon \sim 2\sqrt{N} \|\mathbf{Z}\|_{\mathcal{A}} \epsilon^{-\frac{1}{2}}, \quad (11.199)$$

i.e., the two quantities above are equivalent infinitesimals.

As a result, the new sparse metric  $\mathcal{M}^\epsilon(\mathbf{Z})$  bridges the atomic norm and the atomic  $\ell_0$  norm. As  $\epsilon$  approaches  $+\infty$ , it approaches the former which is convex and can be globally computed but suffers from a resolution limit. As  $\epsilon$  approaches 0, it approaches the latter that exploits sparsity to the greatest extent possible and has no resolution limit but cannot be directly computed.

### 11.6.6.2 A locally convergent iterative algorithm

Inserting Eq. (11.197) into Eq. (11.175), we obtain the following optimization problem:

$$\begin{aligned} & \min_{X, u, Z} \log |\mathbf{T}(u) + \epsilon I| + \text{Tr}(X), \\ & \text{subject to } \begin{bmatrix} X & Z^H \\ Z & \mathbf{T}(u) \end{bmatrix} \geq 0 \text{ and } \|Z_\Omega - Y_\Omega\|_F \leq \eta. \end{aligned} \quad (11.200)$$

This problem is nonconvex since the log-det function is nonconvex. In fact,  $\log |\mathbf{T}(u) + \epsilon I|$  is a concave function of  $u$  since  $\log |\mathbf{R}|$  is a concave function of  $\mathbf{R}$  on the positive semidefinite cone [164]. A popular locally convergent approach to the minimization of such a concave + convex function is the majorization-minimization (MM) algorithm (see, e.g., [162]). Let  $u_j$  denote the  $j$ th iterate of the optimization variable  $u$ . Then, at the  $(j+1)$ th iteration we replace  $\ln |\mathbf{T}(u) + \epsilon I|$  by its tangent plane at the current value  $u = u_j$ .

---

<sup>3</sup> $\mathbf{u}_0$  is called a cluster point of a vector-valued function  $\mathbf{u}(x)$  at  $x = x_0$  if there exists a sequence  $\{x_n\}_{n=1}^{+\infty}$ ,  $\lim_{n \rightarrow +\infty} x_n = x_0$ , satisfying  $\lim_{n \rightarrow +\infty} \mathbf{u}(x_n) = \mathbf{u}_0$ .

$$\begin{aligned} & \ln |\mathbf{T}(\mathbf{u}_j) + \epsilon \mathbf{I}| + \text{Tr} \left[ (\mathbf{T}(\mathbf{u}_j) + \epsilon \mathbf{I})^{-1} \mathbf{T}(\mathbf{u} - \mathbf{u}_j) \right] \\ &= \text{Tr} \left[ (\mathbf{T}(\mathbf{u}_j) + \epsilon \mathbf{I})^{-1} \mathbf{T}(\mathbf{u}) \right] + c_j, \end{aligned} \quad (11.201)$$

where  $c_j$  is a constant independent of  $\mathbf{u}$ . As a result, the optimization problem at the  $(j+1)$ th iteration becomes

$$\begin{aligned} & \min_{X, \mathbf{u}, Z} \text{Tr} \left[ (\mathbf{T}(\mathbf{u}_j) + \epsilon \mathbf{I})^{-1} \mathbf{T}(\mathbf{u}) \right] + \text{Tr}(X), \\ & \text{subject to } \begin{bmatrix} X & Z^H \\ Z & \mathbf{T}(\mathbf{u}) \end{bmatrix} \geq 0 \text{ and } \|Z_\Omega - Y_\Omega\|_F \leq \eta. \end{aligned} \quad (11.202)$$

Note that the problem in Eq. (11.202) is an SDP that can be globally solved. Since  $\log |\mathbf{T}(\mathbf{u}) + \epsilon \mathbf{I}|$  is strictly concave in  $\mathbf{u}$ , at each iteration its value decreases by an amount greater than the decrease of its tangent plane. It follows that by iteratively solving Eq. (11.202) the objective function in Eq. (11.200) monotonically decreases and converges to a local minimum.

### 11.6.6.3 Interpretation as RAM

We show in this subsection that Eq. (11.202) is actually a weighted atomic norm minimization problem. To do so, let us define a weighted atomic set as (compare with the original atomic set  $\mathcal{A}$  defined in Eq. 11.179):

$$\mathcal{A}^w = \{\mathbf{a}^w(f, \boldsymbol{\phi}) = w(f)\mathbf{a}(f)\boldsymbol{\phi} : f \in \mathbb{T}, \boldsymbol{\phi} \in \mathbb{C}^{1 \times L}, \|\boldsymbol{\phi}\|_2 = 1\}, \quad (11.203)$$

where  $w(f) \geq 0$  denotes a weighting function. For  $Z \in \mathbb{C}^{N \times L}$ , define its weighted atomic norm as the atomic norm induced by  $\mathcal{A}^w$ :

$$\begin{aligned} \|Z\|_{\mathcal{A}^w} &= \inf_{c_k, f_k, \boldsymbol{\phi}_k} \left\{ \sum_k c_k : Z = \sum_k c_k \mathbf{a}^w(f_k, \boldsymbol{\phi}_k), f_k \in \mathbb{T}, \|\boldsymbol{\phi}\|_2 = 1, c_k > 0 \right\} \\ &= \inf_{f_k, s_k} \left\{ \sum_k \frac{\|s_k\|_2}{w(f_k)} : Z = \sum_k \mathbf{a}(f_k)s_k \right\}. \end{aligned} \quad (11.204)$$

According to the definition above,  $w(f)$  specifies the importance of the atom at  $f$ : the frequency  $f \in \mathbb{T}$  is more likely to be selected if  $w(f)$  is larger. The atomic norm is a special case of the weighted atomic norm for a constant weighting function. Similar to the atomic norm, the proposed weighted atomic norm also admits a semidefinite formulation for an appropriate weighting function, which is stated in the following theorem.

**Theorem 11.19 ([86])** Suppose that  $w(f) = \frac{1}{\sqrt{\mathbf{a}(f)^H \mathbf{W} \mathbf{a}(f)}} \geq 0$  with  $\mathbf{W} \in \mathbb{C}^{N \times N}$ . Then,

$$\begin{aligned} \|Z\|_{\mathcal{A}^w} &= \min_{X, \mathbf{u}} \frac{\sqrt{N}}{2} \text{Tr}(\mathbf{W} \mathbf{T}(\mathbf{u})) + \frac{1}{2\sqrt{N}} \text{Tr}(X), \\ &\text{subject to } \begin{bmatrix} X & Z^H \\ Z & \mathbf{T}(\mathbf{u}) \end{bmatrix} \geq 0. \end{aligned} \quad (11.205)$$

Let  $\mathbf{W}_j = \frac{1}{N}(\mathbf{T}(\mathbf{u}_j) + \epsilon\mathbf{I})^{-1}$  and  $w_j(f) = \frac{1}{\sqrt{\mathbf{a}(f)^H \mathbf{W}_j \mathbf{a}(f)}}$ . It follows from [Theorem 11.19](#) that the optimization problem in Eq. (11.202) can be exactly written as the following weighted atomic norm minimization problem:

$$\min_{\mathbf{Z}} \|\mathbf{Z}\|_{\mathcal{A}^{w_j}}, \text{ subject to } \|\mathbf{Z}_{\Omega} - \mathbf{Y}_{\Omega}\|_F \leq \eta. \quad (11.206)$$

As a result, the whole iterative algorithm is referred to as reweighted atomic-norm minimization (RAM). Note that the weighting function is updated based on the latest solution  $\mathbf{u}$ . If we let  $w_0(f)$  be constant or equivalently,  $\mathbf{u}_0 = \mathbf{0}$ , such that no preference of the atoms is specified at the first iteration, then the first iteration coincides with ANM. From the second iteration on, the preference is defined by the weighting function  $w_j(f)$  given above. Note that  $w_j^2(f)$  is nothing but the power spectrum of Capon's beamformer (interpreting  $\mathbf{T}(\mathbf{u}_j)$  as the noiseless data covariance and  $\epsilon$  as the noise variance); also note that similar weighting functions have also appeared in sparse optimization methods in the discrete setting (see, e.g., [60, 65, 160]). The reweighting strategy makes the frequencies around those produced by the current iteration more preferable at the next iteration and thus enhances sparsity. At the same time, the weighting results in resolving finer details of the frequency spectrum in those areas and therefore enhances resolution. In a practical implementation of RAM, we can start with the standard ANM, which corresponds to the case of  $\epsilon \rightarrow +\infty$  (by [Theorem 11.18](#)), and then gradually decrease  $\epsilon$  during the iterations.

### 11.6.7 CONNECTIONS BETWEEN ANM AND GLS

We have extended both the atomic norm and the GLS methods from the single to the multiple snapshot case. These two methods were shown in [Section 11.6.3.7](#) to be strongly connected to each other in the single snapshot case, so it is natural to ask whether they are also connected in the multiple snapshot case. We answer this question in this subsection following [84]. In particular, for a small number of snapshots the GLS optimization problem is shown to be equivalent to ANM as if there were no noise, whereas for a large number of snapshots it is equivalent to a weighted ANM. Similar results can also be proved for their discrete versions, viz.  $\ell_{2,1}$  optimization and SPICE.

#### 11.6.7.1 The case of $L < M$

We first consider the ULA case where the GLS optimization problem is given by Eq. (11.165):

$$\min_{X, u} \text{Tr}(X) + \text{Tr}(\mathbf{T}(u)), \text{ subject to } \begin{bmatrix} X & \tilde{\mathbf{Y}}^H \\ \tilde{\mathbf{Y}} & \mathbf{T}(u) \end{bmatrix} \geq 0, \quad (11.207)$$

where  $\tilde{\mathbf{Y}} = \frac{1}{L} \mathbf{Y} (\mathbf{Y}^H \mathbf{Y})^{\frac{1}{2}}$ . By comparing Eq. (11.207) and the SDP formulation of the atomic norm in Eq. (11.187), it can be seen that GLS actually computes  $\|\tilde{\mathbf{Y}}\|_{\mathcal{A}}$  (up to a scaling factor).

A similar argument also holds true in the SLA case where the GLS optimization problem is given by Eq. (11.167):

$$\min_{X,u} \text{Tr}(X) + \frac{M}{N} \text{Tr}(\mathbf{T}(u)), \text{ subject to } \begin{bmatrix} X & \tilde{\mathbf{Y}}_\Omega^H \\ \tilde{\mathbf{Y}}_\Omega & \mathbf{\Gamma}_\Omega \mathbf{T}(u) \mathbf{\Gamma}_\Omega^T \end{bmatrix} \geq 0, \quad (11.208)$$

where  $\tilde{\mathbf{R}}_\Omega$  in Eq. (11.167) is replaced here by  $\tilde{\mathbf{Y}}_\Omega = \frac{1}{L} \mathbf{Y}_\Omega (\mathbf{Y}_\Omega^H \mathbf{Y}_\Omega)^{\frac{1}{2}}$  to reduce the dimensionality. Given  $\mathbf{T}(u) \geq 0$  and applying the identity in Eq. (11.158), we have that

$$\begin{aligned} & \text{Tr}\left(\tilde{\mathbf{Y}}_\Omega^H [\mathbf{\Gamma}_\Omega \mathbf{T}(u) \mathbf{\Gamma}_\Omega^T]^{-1} \tilde{\mathbf{Y}}_\Omega\right) \\ &= \sum_{t=1}^L \tilde{\mathbf{Y}}_\Omega^H(t) [\mathbf{\Gamma}_\Omega \mathbf{T}(u) \mathbf{\Gamma}_\Omega^T]^{-1} \tilde{\mathbf{Y}}_\Omega(t) \\ &= \min_{z(t)} \sum_{t=1}^L z(t)^H [\mathbf{T}(u)]^{-1} z(t), \text{ subject to } z_\Omega(t) = \tilde{\mathbf{Y}}_\Omega(t) \\ &= \min_{\mathbf{Z}} \text{Tr}\left(\mathbf{Z}^H [\mathbf{T}(u)]^{-1} \mathbf{Z}\right), \text{ subject to } \mathbf{Z}_\Omega = \tilde{\mathbf{Y}}_\Omega. \end{aligned} \quad (11.209)$$

It follows that

$$\begin{aligned} \text{Eq. (11.208)} &\Leftrightarrow \min_u \text{Tr}\left(\tilde{\mathbf{Y}}_\Omega^H [\mathbf{\Gamma}_\Omega \mathbf{T}(u) \mathbf{\Gamma}_\Omega^T]^{-1} \tilde{\mathbf{Y}}_\Omega\right) + \frac{M}{N} \text{Tr}(\mathbf{T}(u)) \\ &\Leftrightarrow \min_{u,\mathbf{Z}} \text{Tr}\left(\mathbf{Z}^H [\mathbf{T}(u)]^{-1} \mathbf{Z}\right) + \frac{M}{N} \text{Tr}(\mathbf{T}(u)), \text{ subject to } \mathbf{Z}_\Omega = \tilde{\mathbf{Y}}_\Omega \\ &\Leftrightarrow \min_{X,u,\mathbf{Z}} \frac{M}{N} \text{Tr}(X) + \frac{M}{N} \text{Tr}(\mathbf{T}(u)), \\ &\text{subject to } \begin{bmatrix} X & \sqrt{\frac{N}{M}} \mathbf{Z}^H \\ \sqrt{\frac{N}{M}} \mathbf{Z} & \mathbf{T}(u) \end{bmatrix} \geq 0 \text{ and } \mathbf{Z}_\Omega = \tilde{\mathbf{Y}}_\Omega. \end{aligned} \quad (11.210)$$

The above SDP is nothing but the following ANM problem (up to a scaling factor):

$$\min_{\mathbf{Z}} \|\mathbf{Z}\|_{\mathcal{A}}, \text{ subject to } \mathbf{Z}_\Omega = \tilde{\mathbf{Y}}_\Omega. \quad (11.211)$$

We have therefore shown that when  $L < M$  the GLS optimization problem is equivalent to certain atomic norm formulations obtained by transforming the observed snapshots. Note that the joint sparsity is preserved in the transformed snapshots in the limiting noiseless case. Therefore, in the absence of noise, by applying the results on the atomic norm, GLS is expected to exactly recover the frequencies under the frequency separation condition. This is true in the ULA case where Theorem 11.15 can be directly applied. However, technically, a similar theoretical guarantee cannot be provided in the SLA case since the assumption on the phases in Theorem 11.16 might not hold true for the transformed source signals.

### 11.6.7.2 The case of $L \geq M$

In this case and for a ULA the GLS optimization problem is given by (see Eq. 11.161):

$$\min_{X,u} \text{Tr}(X) + \text{Tr}\left(\tilde{\mathbf{R}}^{-1}\mathbf{T}(u)\right), \text{ subject to } \begin{bmatrix} X & \tilde{\mathbf{R}}^{\frac{1}{2}} \\ \tilde{\mathbf{R}}^{\frac{1}{2}} & \mathbf{T}(u) \end{bmatrix} \geq 0. \quad (11.212)$$

According to Eq. (11.205), this is nothing but computing the weighted atomic norm  $\left\| \tilde{\mathbf{R}}^{\frac{1}{2}} \right\|_{\mathcal{A}^w}$  (up to a scaling factor), where the weighting function is given by  $w(f) = \frac{1}{\sqrt{\mathbf{a}^H(f)\tilde{\mathbf{R}}^{-1}\mathbf{a}(f)}}$ . Note that  $w^2(f) = \frac{1}{\mathbf{a}^H(f)\tilde{\mathbf{R}}^{-1}\mathbf{a}(f)}$  is the power spectrum of the Capon's beamformer.

For an SLA, the GLS problem is given by (see Eq. 11.166):

$$\min_{X,u} \text{Tr}(X) + \text{Tr}\left(\mathbf{\Gamma}_\Omega^T \tilde{\mathbf{R}}_\Omega^{-1} \mathbf{\Gamma}_\Omega \mathbf{T}(u)\right), \text{ subject to } \begin{bmatrix} X & \tilde{\mathbf{R}}_\Omega^{\frac{1}{2}} \\ \tilde{\mathbf{R}}_\Omega^{\frac{1}{2}} & \mathbf{\Gamma}_\Omega \mathbf{T}(u) \mathbf{\Gamma}_\Omega^T \end{bmatrix} \geq 0. \quad (11.213)$$

By arguments similar to those in the preceding subsection we have that

$$\begin{aligned} \text{Eq. (11.213)} &\Leftrightarrow \min_{X,u,Z} \text{Tr}(X) + \text{Tr}\left(\mathbf{\Gamma}_\Omega^T \tilde{\mathbf{R}}_\Omega^{-1} \mathbf{\Gamma}_\Omega \mathbf{T}(u)\right), \text{ subject to } \begin{bmatrix} X & \mathbf{Z}_\Omega^H \\ \mathbf{Z}_\Omega & \mathbf{T}(u) \end{bmatrix} \geq 0 \\ &\quad \text{and } \mathbf{Z}_\Omega = \tilde{\mathbf{R}}_\Omega^{\frac{1}{2}} \\ &\Leftrightarrow \min_Z \|Z\|_{\mathcal{A}^w}, \text{ subject to } \mathbf{Z}_\Omega = \tilde{\mathbf{R}}_\Omega^{\frac{1}{2}}. \end{aligned} \quad (11.214)$$

In Eq. (11.214), the weighting function of the weighted atomic norm is given by (up to a scaling factor)

$$w(f) = \frac{1}{\sqrt{\mathbf{a}^H(f)\mathbf{\Gamma}_\Omega^T \tilde{\mathbf{R}}_\Omega^{-1} \mathbf{\Gamma}_\Omega \mathbf{a}(f)}} = \frac{1}{\sqrt{\mathbf{a}_\Omega^H(f)\tilde{\mathbf{R}}_\Omega^{-1}\mathbf{a}_\Omega(f)}}. \quad (11.215)$$

Therefore, here too  $w^2(f)$  is the power spectrum of the Capon's beamformer.

We have shown above that for a sufficiently large number of snapshots ( $L \geq M$ ) GLS corresponds to a weighted atomic norm in which the weighting function is given by the square root of the Capon's spectrum. While the standard atomic norm method suffers from a “resolution limit” of  $\frac{2.52}{N}$ , the above analysis shows that this limit can actually be overcome by using the weighted atomic norm method: indeed GLS is a consistent method and a large-snapshot realization of the MLE. However, it is worth noting that the standard atomic norm method can be applied to general source signals and its “resolution limit” is given by a worst case analysis, whereas the statistical properties of GLS are obtained under stronger statistical assumptions on the sources and its performance can degrade if these assumptions are not satisfied. The presented connections between GLS and ANM also imply that GLS is generally robust to source correlations, like ANM, though its power estimates can be biased [84].

### 11.6.8 COMPUTATIONAL ISSUES AND SOLUTIONS

We have presented several gridless sparse methods for DOA/frequency estimation from multiple snapshots. Typically, these methods require computing the solution of an SDP. While several off-the-shelf solvers exist for solving SDP, they generally

have high computational complexity. As an example, SDPT3 is an interior-point-based method which has the computational complexity of  $O(n_1^2 n_2^{2.5})$ , where  $n_1$  denotes the number of variables and  $n_2 \times n_2$  is the dimension of the PSD matrix of the SDP [165, 166]. To take a look at the computational complexity of the gridless sparse methods, we consider the atomic norm in Eq. (11.187) as an example. Given  $\mathbf{Z}$  it can be seen that  $n_1 = N + L^2$  and  $n_2 = N + L$ . So the computational complexity of computing the atomic norm can be rather large, viz.  $O((N+L^2)^2(N+L)^{2.5})$ . In this subsection we present strategies for accelerating the computation.

#### 11.6.8.1 Dimensionality reduction

We show in this subsection that a similar dimensionality reduction technique as introduced in Section 11.4.3.3 can be applied to the atomic norm and the weighted atomic norm methods in the case when the number of snapshots  $L$  is large. The technique was firstly proposed in [86] for the gridless setting studied here and it was extended to the discrete case in Section 11.4.3.3 (note that similar techniques were also reported in [167, 168] later on). It is also worth noting that a similar dimensionality reduction is not required by GLS since GLS is covariance-based and all the information in the data snapshots  $\mathbf{Y}_\Omega$  is encoded in the sample covariance matrix  $\tilde{\mathbf{R}}_\Omega = \frac{1}{L} \mathbf{Y}_\Omega^H \mathbf{Y}_\Omega$  whose dimension does not increase with  $L$ . Since it has been shown that GLS and the (weighted) atomic norm are strongly connected, we may naturally wonder if the dimensionality of the atomic norm can be similarly reduced. An affirmative answer is provided in the following result.

**Theorem 11.20** ([86]) *Consider the three ANM problems resulting from Eqs. (11.175)–(11.177) which, respectively, are given by:*

$$\begin{aligned} & \min_{X, u, Z} \text{Tr}(X) + \text{Tr}(T(u)), \\ & \text{subject to } \begin{bmatrix} X & Z^H \\ Z & T(u) \end{bmatrix} \geq \mathbf{0} \text{ and } \|Z_\Omega - Y_\Omega\|_F \leq \eta, \end{aligned} \quad (11.216)$$

$$\begin{aligned} & \min_{X, u, Z} \lambda' \text{Tr}(X) + \lambda' \text{Tr}(T(u)) + \|Z_\Omega - Y_\Omega\|_F^2, \\ & \text{subject to } \begin{bmatrix} X & Z^H \\ Z & T(u) \end{bmatrix} \geq \mathbf{0}, \end{aligned} \quad (11.217)$$

$$\begin{aligned} & \min_{X, u, Z} \text{Tr}(X) + \text{Tr}(T(u)), \\ & \text{subject to } \begin{bmatrix} X & Z^H \\ Z & T(u) \end{bmatrix} \geq \mathbf{0} \text{ and } Z_\Omega = Y_\Omega. \end{aligned} \quad (11.218)$$

Let  $\tilde{Y}_\Omega$  be any matrix satisfying  $\tilde{Y}_\Omega \tilde{Y}_\Omega^H = Y_\Omega Y_\Omega^H$ , such as the  $M \times M$  matrix  $(Y_\Omega Y_\Omega^H)^{\frac{1}{2}}$ . If we replace  $Y_\Omega$  by  $\tilde{Y}_\Omega$  in Eqs. (11.216)–(11.218) and correspondingly change the dimensions of  $Z$  and  $X$ , then the solution  $u$  before and after the replacement is the same. Moreover, if we can find a matrix  $Q$  satisfying  $Q^H Q = I$  and

$\tilde{\mathbf{Y}}_\Omega = \mathbf{Y}_\Omega \mathbf{Q}$  and if  $(\hat{\mathbf{X}}, \hat{\mathbf{Z}}, \hat{\mathbf{u}})$  is the solution after the replacement, then the solution to the original problems is given by  $(\hat{\mathbf{Q}}\hat{\mathbf{X}}\hat{\mathbf{Q}}^H, \hat{\mathbf{Z}}\hat{\mathbf{Q}}^H, \hat{\mathbf{u}})$ .

**Corollary 11.2 ([86])** Theorem 11.20 also holds true if the atomic norm is replaced by the weighted atomic norm.

The dimensionality reduction technique provided by Theorem 11.20 enables us to reduce the number of snapshots from  $L$  to  $M$  and yet obtain the same solution  $\mathbf{u}$ , from which the frequencies and the powers can be retrieved using the Vandermonde decomposition. Therefore, it follows from Theorem 11.20 that for ANM, like for GLS, the information in  $\mathbf{Y}_\Omega$  is preserved in the sample covariance matrix  $\tilde{\mathbf{R}}_\Omega$ . It is interesting to note that the above property even holds true in the presence of coherent sources, while we might expect that DOA estimation from  $\tilde{\mathbf{R}}_\Omega$  could fail in such a case (consider MUSIC as an example).

### 11.6.8.2 Alternating direction method of multipliers (ADMM)

A reasonably fast algorithm for SDPs is based on the ADMM [57, 83, 86, 133], which is a first-order algorithm that guarantees global optimality. To derive the ADMM algorithm, we consider Eq. (11.216) as an example. Define

$$\mathcal{S} = \{\mathbf{Z} : \|\mathbf{Z}_\Omega - \mathbf{Y}_\Omega\|_{\text{F}} \leq \eta\}. \quad (11.219)$$

Then, Eq. (11.216) can be re-written as:

$$\begin{aligned} & \min_{\mathbf{X}, \mathbf{u}, \mathbf{Z} \in \mathcal{S}, \mathbf{Q} \geq 0} \text{Tr}(\mathbf{X}) + \text{Tr}(\mathbf{T}(\mathbf{u})), \\ & \text{subject to } \mathbf{Q} = \begin{bmatrix} \mathbf{X} & \mathbf{Z}^H \\ \mathbf{Z} & \mathbf{T}(\mathbf{u}) \end{bmatrix}. \end{aligned} \quad (11.220)$$

We will derive the algorithm following the routine of ADMM by taking  $(\mathbf{X}, \mathbf{u}, \mathbf{Z})$  and  $\mathbf{Q}$  as the two variables. We introduce  $\Lambda$  as the Lagrangian multiplier and write the augmented Lagrange function for Eq. (11.220) as

$$\begin{aligned} \mathcal{L}(\mathbf{X}, \mathbf{u}, \mathbf{Z}, \mathbf{Q}, \Lambda) &= \text{Tr}(\mathbf{X}) + \text{Tr}(\mathbf{T}(\mathbf{u})) + \text{Tr} \left[ \left( \mathbf{Q} - \begin{bmatrix} \mathbf{X} & \mathbf{Z}^H \\ \mathbf{Z} & \mathbf{T}(\mathbf{u}) \end{bmatrix} \right) \Lambda \right] \\ &+ \frac{\beta}{2} \left\| \mathbf{Q} - \begin{bmatrix} \mathbf{X} & \mathbf{Z}^H \\ \mathbf{Z} & \mathbf{T}(\mathbf{u}) \end{bmatrix} \right\|_{\text{F}}^2, \end{aligned} \quad (11.221)$$

where  $\beta > 0$  is a penalty parameter set according to [57]. The algorithm is implemented by iteratively updating  $(\mathbf{X}, \mathbf{u}, \mathbf{Z})$ ,  $\mathbf{Q}$  and  $\Lambda$  as:

$$(\mathbf{X}, \mathbf{u}, \mathbf{Z}) \leftarrow \arg \min_{\mathbf{X}, \mathbf{u}, \mathbf{Z} \in \mathcal{S}} \mathcal{L}(\mathbf{X}, \mathbf{u}, \mathbf{Z}, \mathbf{Q}, \Lambda), \quad (11.222)$$

$$\mathbf{Q} \leftarrow \arg \min_{\mathbf{Q} \geq 0} \mathcal{L}(\mathbf{X}, \mathbf{u}, \mathbf{Z}, \mathbf{Q}, \Lambda), \quad (11.223)$$

$$\boldsymbol{\Lambda} \leftarrow \boldsymbol{\Lambda} + \beta \left( \mathcal{Q} - \begin{bmatrix} \mathbf{X} & \mathbf{Z}^H \\ \mathbf{Z} & \mathbf{T}(\mathbf{u}) \end{bmatrix} \right). \quad (11.224)$$

Note that  $\mathcal{L}$  in Eq. (11.222) is separable and quadratic in  $\mathbf{X}$ ,  $\mathbf{u}$ , and  $\mathbf{Z}$ . Consequently, these variables can be separately solved for in closed form (note that  $\mathbf{Z}$  can be obtained by firstly solving for  $\mathbf{Z}$  without the set constraint and then projecting the result onto  $\mathcal{S}$ ). In Eq. (11.223),  $\mathcal{Q}$  can be similarly obtained by firstly solving for  $\mathcal{Q}$  without considering the constraint and then projecting the result onto the semidefinite cone, which can be accomplished by forming the eigen-decomposition and setting the negative eigenvalues to zero. The ADMM algorithm has a per-iteration computational complexity of  $O((N+L)^3)$  due to the eigen-decomposition. In the case of  $L > M$ , this complexity can be reduced to  $O((N+M)^3)$  by applying the dimensionality reduction technique presented in the preceding subsection. Although the ADMM may converge slowly to an extremely accurate solution, moderate accuracy is typically sufficient in practical applications [57].

## 11.7 FUTURE RESEARCH CHALLENGES

In this section we highlight several research challenges that should be investigated in future studies.

- **Improving speed and accuracy:** This is a permanent goal of the research on DOA estimation. As compared to most conventional approaches, in general, the sparse methods may have improved DOA estimation accuracy, especially in difficult scenarios, e.g., the cases with no prior knowledge on the source number, few snapshots and coherent sources. But they are more computationally expensive, which is especially true for the recent gridless sparse methods that typically need to solve an SDP. So efficient SDP solvers should be studied in future research, especially when the array size is large. Note that the computation of  $\ell_1$  optimization has been greatly accelerated during the past decade with the development of compressed sensing.

Furthermore, since the convex sparse methods may suffer from certain resolution limits, it will be of interest to study methods that can directly work with  $\ell_0$  norms to enhance the resolution. In the case of ULA and SLA, the  $\ell_0$  optimization corresponds to matrix rank minimization. Therefore, it is possible to apply the recent nonconvex optimization techniques for rank minimization to DOA estimation (see, e.g., [169–174]). Recent progresses in this direction have been made in [175, 176].

- **Automatic model order selection:** Different from the conventional subspace-based methods, the sparse optimization methods usually do not require a prior knowledge of the source number (a.k.a. the model order). But this does not mean that the sparse optimization methods are able to accurately estimate the source number. Instead, small spurious sources are usually present in the obtained power

spectrum. Therefore, it is of great interest to study how the model order can be automatically estimated within or after the use of the sparse methods. Some results in this direction can be found in [133, 149, 177]. In [133] a parameter refining technique is also introduced once the model order is obtained.

- **Gridless sparse methods for arbitrary array geometry:** The grid-based sparse methods can be applied to any array geometry but they suffer from grid mismatch or other problems. In contrast to this, the gridless sparse methods completely bypass the grid selection problem by utilizing the special Hankel/Toeplitz structure of the sampled data or the data covariance matrix in the case of ULAs and SLAs. For a general array geometry, however, such structures do not exist anymore and extension of the gridless sparse methods to general arrays should be studied in future. Recent results in this direction can be found in [178, 179].
- **Gridless sparse parameter estimation and continuous compressed sensing:** Following the line of the previous discussion, it would be of great interest to extend the existing gridless sparse methods for DOA estimation to general parameter estimation problems. Note that a similar data model, as used in DOA estimation, can be formulated for a rather general parameter estimation problem:

$$\mathbf{y} = \sum_{k=1}^K \mathbf{a}(\theta_k) x_k + \mathbf{e}, \quad (11.225)$$

where  $\mathbf{y}$  is the data vector,  $\mathbf{a}(\theta)$  is a given function of the continuous parameter  $\theta$ ,  $x_k$  are weight coefficients and  $\mathbf{e}$  denotes noise. Moreover, to guarantee that the parameters are identifiable as well as to simplify the model in Eq. (11.225), it is natural to assume that “order”  $K$  is small and thus sparsity concept can be introduced as well. But due to the absence of special Hankel/Toeplitz structures, it would be challenging to develop gridless methods for Eq. (11.225). Note that the estimation of  $\theta_k$  and  $x_k$ ,  $k = 1, \dots, K$  from  $\mathbf{y}$  based on the data model in Eq. (11.225) is also referred to as continuous or infinite-dimensional compressed sensing, which extends compressed sensing from the discrete to the continuous setting [130, 157, 180].

---

## 11.8 CONCLUSIONS

In this article, we provided an overview of the sparse DOA estimation techniques. Two key differences between sparse representation and DOA estimation were pointed out: (1) discrete system versus continuous parameters and (2) single versus multiple snapshots. Based on how the first difference is dealt with, the sparse methods were classified and discussed in three categories, namely on-grid, off-grid, and gridless. The second difference can be tackled by exploiting the temporal redundancy of the snapshots. We explained that while the on-grid and off-grid sparse methods can be applied to arbitrary array geometries, they may suffer from grid mismatch, weak theoretical guarantees, etc. These drawbacks can be eliminated by using the gridless sparse methods which, however, can only be applied to ULAs and SLAs. We also highlighted some challenging problems that should be studied in future

research. Note that these sparse methods have diverse applications to many fields and the future work also includes performance comparisons of these methods for each specific application. Depending on data qualities and quantities, one or more of these methods may be favored in one application but not another.

---

## ACKNOWLEDGMENTS

Z. Yang and L. Xie were supported by the Ministry of Education, Republic of Singapore, under grant AcRF TIER 1 RG78/15. Z. Yang was also supported by the National Natural Science Foundation of China, under grant 61603187 and by the Natural Science Foundation of Jiangsu Province, China, under grant BK20160845. J. Li was supported by the National Science Foundation (NSF) under grant CCF-1218388. P. Stoica was supported by a Swedish Research Council (VR) grant.

---

## REFERENCES

- [1] J. Capon, High-resolution frequency-wavenumber spectrum analysis, *Proc. IEEE* 57 (8) (1969) 1408–1418.
- [2] V.F. Pisarenko, The retrieval of harmonics from a covariance function, *Geophys. J. Int.* 33 (3) (1973) 347–366.
- [3] R. Schmidt, A Signal Subspace Approach to Multiple Emitter Location Spectral Estimation, Ph.D. thesis, Stanford University, 1981.
- [4] R.O. Schmidt, Multiple emitter location and signal parameter estimation, *IEEE Trans. Antennas Propag.* 34 (3) (1986) 276–280.
- [5] A. Paulraj, R. Roy, T. Kailath, A subspace rotation approach to signal parameter estimation, *Proc. IEEE* 74 (7) (1986) 1044–1046.
- [6] R. Roy, T. Kailath, ESPRIT-estimation of signal parameters via rotational invariance techniques, *IEEE Trans. Acoust. Speech Signal Process.* 37 (7) (1989) 984–995.
- [7] A. Barabell, Improving the resolution performance of eigenstructure-based direction-finding algorithms, in: *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 8, 1983, pp. 336–339.
- [8] H. Krim, M. Viberg, Two decades of array signal processing research: the parametric approach, *IEEE Signal Process. Mag.* 13 (4) (1996) 67–94.
- [9] P. Stoica, R.L. Moses, *Spectral Analysis of Signals*, Pearson/Prentice Hall, Upper Saddle River, NJ, 2005.
- [10] A. Zoubir, M. Viberg, R. Chellappa, S. Theodoridis, *Academic Press Library in Signal Processing Volume 3: Array and Statistical Signal Processing*, Academic Press, New York, NY, USA, 2014.
- [11] S. Chen, D. Donoho, M. Saunders, Atomic decomposition by basis pursuit, *SIAM Rev.* 43 (1) (2001) 129–159.
- [12] D.L. Donoho, M. Elad, Optimally sparse representation in general (nonorthogonal) dictionaries via  $\ell^1$  minimization, *Proc. Natl. Acad. Sci.* 100 (5) (2003) 2197–2202.
- [13] E.J. Candès, J. Romberg, T. Tao, Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information, *IEEE Trans. Inf. Theory* 52 (2) (2006) 489–509.
- [14] D. Donoho, Compressed sensing, *IEEE Trans. Inf. Theory* 52 (4) (2006) 1289–1306.
- [15] R.G. Baraniuk, Compressive sensing, *IEEE Signal Process. Mag.* 24 (4) (2007).

- [16] P.-J. Chung, M. Viberg, J. Yu, DOA estimation methods and algorithms, in: Academic Press Library in Signal Processing Volume 3: Array and Statistical Signal Processing Academic Press, New York, NY, USA, 2014, pp. 599–650.
- [17] Y. Bresler, A. Macovski, On the number of signals resolvable by a uniform linear array, *IEEE Trans. Acoust. Speech Signal Process.* 34 (6) (1986) 1361–1375.
- [18] M. Wax, I. Ziskind, On unique localization of multiple sources by passive sensor arrays, *IEEE Trans. Acoust. Speech Signal Process.* 37 (7) (1989) 996–1000.
- [19] A. Nehorai, D. Starer, P. Stoica, Direction-of-arrival estimation in applications with multipath and few snapshots, *Circ. Syst. Signal Process.* 10 (3) (1991) 327–342.
- [20] J. Chen, X. Huo, Theoretical results on sparse representations of multiple-measurement vectors, *IEEE Trans. Signal Process.* 54 (12) (2006) 4634–4643.
- [21] M.E. Davies, Y.C. Eldar, Rank awareness in joint sparse recovery, *IEEE Trans. Inf. Theory* 58 (2) (2012) 1135–1146.
- [22] Z. Yang, L. Xie, Exact joint sparse frequency recovery via optimization methods, *IEEE Trans. Signal Process.* 64 (19) (2016) 5145–5157.
- [23] J.B. Kruskal, Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics, *Linear Algebra Appl.* 18 (2) (1977) 95–138.
- [24] R.G. Baraniuk, E. Candès, R. Nowak, M. Vetterli, Compressive sampling, *IEEE Signal Process. Mag.* 25 (2) (2008) 12–13.
- [25] R. Chartrand, R.G. Baraniuk, Y.C. Eldar, M.A. Figueiredo, J. Tanner, Introduction to the issue on compressive sensing, *IEEE J. Sel. Top. Signal Process.* 2 (4) (2010) 241–243.
- [26] R.G. Baraniuk, E. Candes, M. Elad, Y. Ma, Applications of sparse representation and compressive sensing [scanning the issue], *Proc. IEEE* 98 (6) (2010) 906–909.
- [27] J.-L. Starck, J. Fadili, M. Elad, R.D. Nowak, P. Tsakalides, Introduction to the issue on Adaptive sparse representation of data and applications in signal and image processing, *IEEE J. Sel. Top. Signal Process.* 5 (5) (2011) 893–895.
- [28] I.F. Gorodnitsky, B.D. Rao, Sparse signal reconstruction from limited data using FOCUSS: a re-weighted minimum norm algorithm, *IEEE Trans. Signal Process.* 45 (3) (1997) 600–616.
- [29] B.D. Rao, K. Kreutz-Delgado, An affine scaling methodology for best basis selection, *IEEE Trans. Signal Process.* 47 (1) (1999) 187–200.
- [30] B.D. Rao, K. Engan, S.F. Cotter, J. Palmer, K. Kreutz-Delgado, Subset selection in noise based on diversity measure minimization, *IEEE Trans. Signal Process.* 51 (3) (2003) 760–770.
- [31] S. Foucart, M. Lai, Sparsest solutions of underdetermined linear systems via  $\ell_q$ -minimization for  $0 < q < 1$ , *Appl. Comput. Harmonic Anal.* 26 (3) (2009) 395–407.
- [32] R. Chartrand, Nonconvex compressed sensing and error correction, in: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 3, 2007, pp. 889–892.
- [33] R. Chartrand, Exact reconstruction of sparse signals via nonconvex minimization, *IEEE Signal Process. Lett.* 14 (10) (2007) 707–710.
- [34] X. Tan, W. Roberts, J. Li, P. Stoica, Sparse learning via iterative minimization with application to MIMO radar imaging, *IEEE Trans. Signal Process.* 59 (3) (2011) 1088–1101.
- [35] Y.C. Pati, R. Rezaifar, P. Krishnaprasad, Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition, in: *1993 Conference*

- Record of The Twenty-Seventh Asilomar Conference on Signals, Systems and Computers, 1993, pp. 40–44.
- [36] D.L. Donoho, Y. Tsaig, I. Drori, J.-L. Starck, Sparse solution of underdetermined systems of linear equations by stagewise orthogonal matching pursuit, *IEEE Trans. Inf. Theory* 58 (2) (2012) 1094–1121.
  - [37] J. Tropp, A. Gilbert, Signal recovery from random measurements via orthogonal matching pursuit, *IEEE Trans. Inf. Theory* 53 (12) (2007) 4655–4666.
  - [38] D. Needell, J. Tropp, CoSaMP: iterative signal recovery from incomplete and inaccurate samples, *Appl. Comput. Harmonic Anal.* 26 (3) (2009) 301–321.
  - [39] W. Dai, O. Milenkovic, Subspace pursuit for compressive sensing signal reconstruction, *IEEE Trans. Inf. Theory* 55 (5) (2009) 2230–2249.
  - [40] M.A. Davenport, M.B. Wakin, Analysis of orthogonal matching pursuit using the restricted isometry property, *IEEE Trans. Inf. Theory* 56 (9) (2010) 4395–4401.
  - [41] T. Blumensath, M. Davies, Iterative hard thresholding for compressed sensing, *Appl. Comput. Harmonic Anal.* 27 (3) (2009) 265–274.
  - [42] J. Tropp, S. Wright, Computational methods for sparse solution of linear inverse problems, *Proc. IEEE* 98 (6) (2010) 948–958.
  - [43] J.F. Claerbout, F. Muir, Robust modeling with erratic data, *Geophysics* 38 (5) (1973) 826–844.
  - [44] E. Candès, Compressive sampling, in: *Proceedings of the International Congress of Mathematicians*, vol. 3, 2006, pp. 1433–1452.
  - [45] E. Candès, The restricted isometry property and its implications for compressed sensing, *C. R. Math.* 346 (9–10) (2008) 589–592.
  - [46] S. Foucart, Sparse recovery algorithms: sufficient conditions in terms of restricted isometry constants, in: *Approximation Theory XIII: San Antonio 2010*, Springer, New York, 2012, pp. 65–77.
  - [47] T. Cai, L. Wang, G. Xu, New bounds for restricted isometry constants, *IEEE Trans. Inf. Theory* 56 (9) (2010) 4388–4394.
  - [48] R. Tibshirani, Regression shrinkage and selection via the Lasso, *J. R. Stat. Soc. Ser. B* 58 (1) (1996) 267–288.
  - [49] A. Belloni, V. Chernozhukov, L. Wang, Square-root lasso: pivotal recovery of sparse signals via conic programming, *Biometrika* 98 (4) (2011) 791–806.
  - [50] E. Candès, J. Romberg,  $\ell_1$ -magic: recovery of sparse signals via convex programming, 2005, Available at: <http://users.ece.gatech.edu/justin/l1magic/downloads/l1magic.pdf>.
  - [51] S. Kim, K. Koh, M. Lustig, S. Boyd, D. Gorinevsky, An interior-point method for large-scale  $\ell_1$ -regularized least squares, *IEEE J. Sel. Top. Signal Process.* 1 (4) (2008) 606–617.
  - [52] M. Lustig, D. Donoho, J. Pauly, Sparse MRI: the application of compressed sensing for rapid MR imaging, *Magn. Reson. Med.* 58 (6) (2007) 1182–1195.
  - [53] E.T. Hale, W. Yin, Y. Zhang, A fixed-point continuation method for  $\ell_1$ -regularized minimization with applications to compressed sensing. CAAM TR07-07, Rice University, 2007.
  - [54] Y. Nesterov, Smooth minimization of non-smooth functions, *Math. Program.* 103 (1) (2005) 127–152.
  - [55] S. Becker, J. Bobin, E. Candès, NESTA: a fast and accurate first-order method for sparse recovery, *SIAM J. Imaging Sci.* 4 (1) (2011) 1–39.
  - [56] Z. Yang, C. Zhang, J. Deng, W. Lu, Orthonormal expansion  $\ell_1$ -minimization algorithms for compressed sensing, *IEEE Trans. Signal Process.* 59 (12) (2011) 6285–6290.

- [57] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers, *Found. Trends Mach. Learn.* 3 (1) (2011) 1–122.
- [58] J. Yang, Y. Zhang, Alternating direction algorithms for  $\ell_1$ -problems in compressive sensing, *SIAM J. Sci. Comput.* 33 (1) (2011) 250–278.
- [59] D.P. Wipf, B.D. Rao, Sparse Bayesian learning for basis selection, *IEEE Trans. Signal Process.* 52 (8) (2004) 2153–2164.
- [60] P. Stoica, P. Babu, SPICE and LIKES: two hyperparameter-free methods for sparse-parameter estimation, *Signal Process.* 92 (7) (2012) 1580–1590.
- [61] M. Tipping, Sparse Bayesian learning and the relevance vector machine, *J. Mach. Learn. Res.* 1 (2001) 211–244.
- [62] S. Ji, Y. Xue, L. Carin, Bayesian compressive sensing, *IEEE Trans. Signal Process.* 56 (6) (2008) 2346–2356.
- [63] P. Stoica, P. Babu, J. Li, New method of sparse parameter estimation in separable models and its use for spectral analysis of irregularly sampled data, *IEEE Trans. Signal Process.* 59 (1) (2011) 35–47.
- [64] P. Stoica, P. Babu, J. Li, SPIce: a sparse covariance-based estimation method for array processing, *IEEE Trans. Signal Process.* 59 (2) (2011) 629–638.
- [65] P. Stoica, D. Zachariah, J. Li, Weighted SPIce: a unifying approach for hyperparameter-free sparse estimation, *Digit. Signal Process.* 33 (2014) 1–12.
- [66] G. McLachlan, T. Krishnan, *The EM Algorithm and Extensions*, John Wiley & Sons, New York, 1997.
- [67] S. Babacan, R. Molina, A. Katsaggelos, Bayesian compressive sensing using Laplace priors, *IEEE Trans. Image Process.* 19 (1) (2010) 53–63.
- [68] S.F. Cotter, B.D. Rao, K. Engan, K. Kreutz-Delgado, Sparse solutions to linear inverse problems with multiple measurement vectors, *IEEE Trans. Signal Process.* 53 (7) (2005) 2477–2488.
- [69] D. Malioutov, M. Cetin, A.S. Willsky, A sparse signal reconstruction perspective for source localization with sensor arrays, *IEEE Trans. Signal Process.* 53 (8) (2005) 3010–3022.
- [70] M. Fornasier, H. Rauhut, Recovery algorithms for vector-valued data with joint sparsity constraints, *SIAM J. Numer. Anal.* 46 (2) (2008) 577–613.
- [71] M. Mishali, Y.C. Eldar, Reduce and boost: recovering arbitrary sets of jointly sparse vectors, *IEEE Trans. Signal Process.* 56 (10) (2008) 4692–4702.
- [72] R. Gribonval, H. Rauhut, K. Schnass, P. Vandergheynst, Atoms of all channels, unite! Average case analysis of multi-channel sparse recovery using greedy algorithms, *J. Fourier Anal. Appl.* 14 (5–6) (2008) 655–687.
- [73] M. Kowalski, Sparse regression using mixed norms, *Appl. Comput. Harmonic Anal.* 27 (3) (2009) 303–324.
- [74] S. Ji, D. Dunson, L. Carin, Multitask compressive sensing, *IEEE Trans. Signal Process.* 57 (1) (2009) 92–106.
- [75] Y.C. Eldar, M. Mishali, Robust recovery of signals from a structured union of subspaces, *IEEE Trans. Inf. Theory* 55 (11) (2009) 5302–5316.
- [76] Y. Eldar, H. Rauhut, Average case analysis of multichannel sparse recovery using convex relaxation, *IEEE Trans. Inf. Theory* 56 (1) (2010) 505–519.
- [77] M. Hyder, K. Mahata, Direction-of-arrival estimation using a mixed  $\ell_{2,0}$  norm approximation, *IEEE Trans. Signal Process.* 58 (9) (2010) 4646–4655.

- [78] E. Van Den Berg, M. Friedlander, Theoretical and empirical results for recovery from multiple measurements., *IEEE Trans. Inf. Theory* 56 (5) (2010) 2516–2527.
- [79] J.M. Kim, O.K. Lee, J.C. Ye, Compressive MUSIC: revisiting the link between compressive sensing and array signal processing, *IEEE Trans. Inf. Theory* 58 (1) (2012) 278–301.
- [80] K. Lee, Y. Bresler, M. Junge, Subspace methods for joint sparse recovery, *IEEE Trans. Inf. Theory* 58 (6) (2012) 3613–3641.
- [81] Y. Chi, L. Scharf, A. Pezeshki, A. Calderbank, Sensitivity to basis mismatch in compressed sensing, *IEEE Trans. Signal Process.* 59 (5) (2011) 2182–2195.
- [82] D. Chae, P. Sadeghi, R. Kennedy, Effects of basis-mismatch in compressive sampling of continuous sinusoidal signals, in: 2nd IEEE International Conference on Future Computer and Communication (ICFCC), vol. 2, 2010, pp. 739–743.
- [83] B.N. Bhaskar, G. Tang, B. Recht, Atomic norm denoising with applications to line spectral estimation, *IEEE Trans. Signal Process.* 61 (23) (2013) 5987–5999.
- [84] Z. Yang, L. Xie, On gridless sparse methods for multi-snapshot direction of arrival estimation, *Circuits Syst. Signal Process.* 36 (8) (2017) 3370–3384.
- [85] Y. Li, Y. Chi, Off-the-grid line spectrum denoising and estimation with multiple measurement vectors, *IEEE Trans. Signal Process.* 64 (5) (2016) 1257–1269.
- [86] Z. Yang, L. Xie, Enhancing sparsity and resolution via reweighted atomic norm minimization, *IEEE Trans. Signal Process.* 64 (4) (2016) 995–1006.
- [87] B. Ottersten, P. Stoica, R. Roy, Covariance matching estimation techniques for array signal processing applications, *Digit. Signal Process.* 8 (3) (1998) 185–210.
- [88] T. Anderson, *Multivariate Statistical Analysis*, Wiley and Sons, New York, NY, 1984.
- [89] H. Li, P. Stoica, J. Li, Computationally efficient maximum likelihood estimation of structured covariance matrices, *IEEE Trans. Signal Process.* 47 (5) (1999) 1314–1323.
- [90] C. Rojas, D. Katselis, H. Hjalmarsson, A note on the SPICE method, *IEEE Trans. Signal Process.* 61 (18) (2013) 4545–4551.
- [91] P. Babu, P. Stoica, Connection between SPICE and Square-Root LASSO for sparse parameter estimation, *Signal Process.* 95 (2014) 10–14.
- [92] Z. Yang, L. Xie, C. Zhang, A discretization-free sparse and parametric approach for linear array signal processing, *IEEE Trans. Signal Process.* 62 (19) (2014) 4959–4973.
- [93] D.P. Wipf, B.D. Rao, An empirical Bayesian strategy for solving the simultaneous sparse approximation problem, *IEEE Trans. Signal Process.* 55 (7) (2007) 3704–3716.
- [94] Z.-M. Liu, Z.-T. Huang, Y.-Y. Zhou, An efficient maximum likelihood method for direction-of-arrival estimation via sparse Bayesian learning, *IEEE Trans. Wirel. Commun.* 11 (10) (2012) 1–11.
- [95] M. Carlin, P. Rocca, G. Oliveri, F. Viani, A. Massa, Directions-of-arrival estimation through Bayesian compressive sensing strategies, *IEEE Trans. Antennas Propag.* 61 (7) (2013) 3828–3838.
- [96] P. Stoica, P. Babu, Sparse estimation of spectral lines: grid selection problems and their solutions, *IEEE Trans. Signal Process.* 60 (2) (2012) 962–967.
- [97] C. Austin, J. Ash, R. Moses, Dynamic dictionary algorithms for model order and parameter estimation, *IEEE Trans. Signal Process.* 61 (20) (2013) 5117–5130.
- [98] M.F. Duarte, R.G. Baraniuk, Spectral compressive sensing, *Appl. Comput. Harmonic Anal.* 35 (1) (2013) 111–129.
- [99] A. Fannjiang, W. Liao, Coherence pattern-guided compressive sensing with unresolved grids, *SIAM J. Imaging Sci.* 5 (1) (2012) 179–202.

- [100] E.J. Candès, C. Fernandez-Granda, Towards a mathematical theory of super-resolution, *Commun. Pure Appl. Math.* 67 (6) (2014) 906–956.
- [101] H. Zhu, G. Leus, G. Giannakis, Sparsity-cognizant total least-squares for perturbed compressive sampling, *IEEE Trans. Signal Process.* 59 (5) (2011) 2002–2016.
- [102] J. Zheng, M. Kaveh, Directions-of-arrival estimation using a sparse spatial spectrum model with uncertainty, in: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 2848–2851.
- [103] Z. Yang, C. Zhang, L. Xie, Robustly stable signal recovery in compressed sensing with structured matrix perturbation, *IEEE Trans. Signal Process.* 60 (9) (2012) 4658–4671.
- [104] Z. Yang, L. Xie, C. Zhang, Off-grid direction of arrival estimation using sparse Bayesian inference, *IEEE Trans. Signal Process.* 61 (1) (2013) 38–43.
- [105] Z. Tan, P. Yang, A. Nehorai, Joint sparse recovery method for compressed sensing with structured dictionary mismatch, *IEEE Trans. Signal Process.* 62 (19) (2014) 4997–5008.
- [106] Y. Zhang, Z. Ye, X. Xu, N. Hu, Off-grid DOA estimation using array covariance matrix and block-sparse Bayesian learning, *Signal Process.* 98 (2014) 197–201.
- [107] M. Lasserre, S. Bidon, O. Besson, F. Le Chevalier, Bayesian sparse Fourier representation of off-grid targets with application to experimental radar data, *Signal Process.* 111 (2015) 261–273.
- [108] W. Si, X. Qu, Z. Qu, Off-grid DOA estimation using alternating block coordinate descent in compressed sensing, *Sensors* 15 (9) (2015) 21099–21113.
- [109] T. Chen, H. Wu, L. Guo, L. Liu, A modified rife algorithm for off-grid DOA estimation based on sparse representations, *Sensors* 15 (11) (2015) 29721–29733.
- [110] L. Wang, L. Zhao, G. Bi, C. Wan, L. Zhang, H. Zhang, Novel wideband DOA estimation based on sparse Bayesian learning with Dirichlet process priors, *IEEE Trans. Signal Process.* 64 (2) (2016) 275–289.
- [111] X. Wu, W.-P. Zhu, J. Yan, Direction of arrival estimation for off-grid signals based on sparse Bayesian learning, *IEEE Sens. J.* 16 (7) (2016) 2004–2016.
- [112] Y. Zhao, L. Zhang, Y. Gu, Array covariance matrix-based sparse Bayesian learning for off-grid direction-of-arrival estimation, *Electron. Lett.* 52 (5) (2016) 401–402.
- [113] G. Han, L. Wan, L. Shu, N. Feng, Two novel DOA estimation approaches for real-time assistant calibration systems in future vehicle industrial, *IEEE Syst. J.* (2015).
- [114] S. Bermhardt, R. Boyer, S. Marcos, P. Larzabal, Compressed sensing with basis mismatch: performance bounds and sparse-based estimator, *IEEE Trans. Signal Process.* 64 (13) (2016) 3483–3494.
- [115] Y. Fei, T. Jian-wu, Z. Qing-jie, Off-grid sparse estimator for air velocity in missing-data case, *J. Aircraft* (2016) 1–10.
- [116] Q. Shen, W. Cui, W. Liu, S. Wu, Y.D. Zhang, M.G. Amin, Underdetermined wideband DOA estimation of off-grid sources employing the difference co-array concept, *Signal Process.* 130 (2017) 299–304.
- [117] J. Yang, G. Liao, J. Li, An efficient off-grid DOA estimation approach for nested array signal processing by using sparse Bayesian learning strategies, *Signal Process.* 128 (2016) 110–122.
- [118] F. Sun, Q. Wu, Y. Sun, G. Ding, P. Lan, An iterative approach for sparse direction-of-arrival estimation in co-prime arrays with off-grid targets, *Digit. Signal Process.* 61 (2017) 35–42.

- [119] D. Shutin, B.H. Fleury, Sparse variational Bayesian SAGE algorithm with application to the estimation of multipath wireless channels, *IEEE Trans. Signal Process.* 59 (8) (2011) 3609–3623.
- [120] D. Shutin, W. Wang, T. Jost, Incremental sparse Bayesian learning for parameter estimation of superimposed signals, in: 10th International Conference on Sampling Theory and Applications, 2013.
- [121] L. Hu, Z. Shi, J. Zhou, Q. Fu, Compressed sensing of complex sinusoids: an approach based on dictionary refinement, *IEEE Trans. Signal Process.* 60 (7) (2012) 3809–3822.
- [122] L. Hu, J. Zhou, Z. Shi, Q. Fu, A fast and accurate reconstruction algorithm for compressed sensing of complex sinusoids, *IEEE Trans. Signal Process.* 61 (22) (2013) 5744–5754.
- [123] J. Fang, J. Li, Y. Shen, H. Li, S. Li, Super-resolution compressed sensing: an iterative reweighted algorithm for joint parameter learning and sparse signal recovery, *IEEE Signal Process. Lett.* 21 (6) (2014) 761–765.
- [124] J. Fang, F. Wang, Y. Shen, H. Li, R. Blum, Super-resolution compressed sensing for line spectral estimation: an iterative reweighted approach, *IEEE Trans. Signal Process.* 64 (18) (2016) 4649–4662.
- [125] C. Carathéodory, L. Fejér, Über den Zusammenhang der Extremen von harmonischen Funktionen mit ihren Koeffizienten und über den Picard-Landau'schen Satz, *Rendiconti del Circolo Matematico di Palermo* (1884–1940) 32 (1) (1911) 218–239.
- [126] L. Gurvits, H. Barnum, Largest separable balls around the maximally mixed bipartite quantum state, *Phys. Rev. A* 66 (6) (2002) 062311.
- [127] R.A. Horn, C.R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, 2012.
- [128] E.J. Candès, C. Fernandez-Granda, Super-resolution from noisy data, *J. Fourier Anal. Appl.* 19 (6) (2013) 1229–1254.
- [129] G. Tang, B.N. Bhaskar, P. Shah, B. Recht, Compressive sensing off the grid, in: 50th IEEE Annual Allerton Conference on Communication, Control, and Computing (Allerton), 2012, pp. 778–785.
- [130] G. Tang, B.N. Bhaskar, P. Shah, B. Recht, Compressed sensing off the grid, *IEEE Trans. Inf. Theory* 59 (11) (2013) 7465–7490.
- [131] G. Tang, B.N. Bhaskar, B. Recht, Near minimax line spectral estimation, *IEEE Trans. Inf. Theory* 61 (1) (2015) 499–512.
- [132] Y. Chen, Y. Chi, Robust spectral compressed sensing via structured matrix completion, *IEEE Trans. Inf. Theory* 60 (10) (2014) 6576–6601.
- [133] Z. Yang, L. Xie, On gridless sparse methods for line spectral estimation from complete and incomplete data, *IEEE Trans. Signal Process.* 63 (12) (2015) 3139–3153.
- [134] J.-M. Azais, Y. De Castro, F. Gamboa, Spike detection from inaccurate samplings, *Appl. Comput. Harmonic Anal.* 38 (2) (2015) 177–195.
- [135] V. Duval, G. Peyré, Exact support recovery for sparse spikes deconvolution, *Found. Comput. Math.* (2015) 1–41.
- [136] C. Fernandez-Granda, Super-resolution of point sources via convex programming, *Inf. Inference* 5 (3) (2016) 251–303.
- [137] J.-F. Cai, X. Qu, W. Xu, G.-B. Ye, Robust recovery of complex exponential signals from random Gaussian projections via low rank Hankel matrix reconstruction, *Appl. Comput. Harmonic Anal.* 41 (2) (2016) 470–490.
- [138] L. Sun, H. Hong, Y. Li, C. Gu, F. Xi, C. Li, X. Zhu, Noncontact vital sign detection based on stepwise atomic norm minimization, *IEEE Signal Process. Lett.* 22 (12) (2015) 2479–2483.

- [139] P. Stoica, G. Tang, Z. Yang, D. Zachariah, Gridless compressive-sensing methods for frequency estimation: Points of tangency and links to basics, in: 22nd European Signal Processing Conference (EUSIPCO), 2014, pp. 1831–1835.
- [140] B. Sun, H. Feng, Z. Zhang, A new approach for heart rate monitoring using photoplethysmography signals contaminated by motion artifacts, in: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 809–813.
- [141] Y. Zhang, X. Hong, Y. Wang, D. Sun, Gridless SPICE applied to parameter estimation of underwater acoustic Frequency Hopping signals, in: 2016 IEEE/OES China Ocean Acoustics (COA), 2016, pp. 1–6.
- [142] Q. Bao, X. Peng, Z. Wang, Y. Lin, W. Hong, DLSLA 3-D SAR imaging based on reweighted gridless sparse recovery method, *IEEE Geosci. Remote Sens. Lett.* 13 (6) (2016) 841–845.
- [143] V. Chandrasekaran, B. Recht, P.A. Parrilo, A.S. Willsky, The convex geometry of linear inverse problems, *Found. Comput. Math.* 12 (6) (2012) 805–849.
- [144] W. Rudin, *Real and Complex Analysis*, Tata McGraw-Hill Education, New York, 1987.
- [145] B.N. Bhaskar, B. Recht, Atomic norm denoising with applications to line spectral estimation, in: 49th IEEE Annual Allerton Conference on Communication, Control, and Computing (Allerton), 2011, pp. 261–268.
- [146] M. Fazel, H. Hindi, S.P. Boyd, A rank minimization heuristic with application to minimum order system approximation, in: American Control Conference, vol. 6, 2001, pp. 4734–4739.
- [147] R. Rochberg, Toeplitz and Hankel operators on the Paley-Wiener space, *Integral Equations Operator Theory* 10 (2) (1987) 187–235.
- [148] Z. He, A. Cichocki, S. Xie, K. Choi, Detecting the number of clusters in  $n$ -way probabilistic clustering, *IEEE Trans. Pattern Anal. Mach. Intell.* 32 (11) (2010) 2006–2021.
- [149] Z. Tan, Y.C. Eldar, A. Nehorai, Direction of arrival estimation using co-prime arrays: a super resolution viewpoint, *IEEE Trans. Signal Process.* 62 (21) (2014) 5565–5576.
- [150] P. Pal, P. Vaidyanathan, A grid-less approach to underdetermined direction of arrival estimation via low rank matrix denoising, *IEEE Signal Process. Lett.* 21 (6) (2014) 737–741.
- [151] D.A. Linebarger, I.H. Sudborough, I.G. Tollis, Difference bases and sparse sensor arrays, *IEEE Trans. Inf. Theory* 39 (2) (1993) 716–721.
- [152] M. Viberg, Introduction to array processing, in: Academic Press Library in Signal Processing Volume 3: Array and Statistical Signal Processing, 2014, pp. 463–502.
- [153] P. Stoica, T. Soderstrom, On reparametrization of loss functions used in estimation and the invariance principle, *Signal Process.* 17 (4) (1989) 383–387.
- [154] P.P. Vaidyanathan, P. Pal, Sparse sensing with co-prime samplers and arrays, *IEEE Trans. Signal Process.* 59 (2) (2011) 573–586.
- [155] C. Fernandez-Granda, Support detection in super-resolution, 2013, arXiv preprint arXiv:1302.3921.
- [156] Z. Yang, L. Xie, P. Stoica, Vandermonde decomposition of multilevel Toeplitz matrices with application to multidimensional super-resolution, *IEEE Trans. Inf. Theory* 62 (6) (2016) 3685–3701.
- [157] Z. Yang, L. Xie, Continuous compressed sensing with a single or multiple measurement vectors, in: IEEE Workshop on Statistical Signal Processing (SSP), 2014, pp. 308–311.
- [158] M.S. Lobo, M. Fazel, S. Boyd, Portfolio optimization with linear and fixed transaction costs, *Ann. Oper. Res.* 152 (1) (2007) 341–365.

- [159] E.J. Candes, M.B. Wakin, S.P. Boyd, Enhancing sparsity by reweighted  $\ell_1$  minimization, *J. Fourier Anal. Appl.* 14 (5–6) (2008) 877–905.
- [160] D. Wipf, S. Nagarajan, Iterative reweighted  $\ell_1$  and  $\ell_2$  methods for finding sparse solutions, *IEEE J. Sel. Top. Signal Process.* 4 (2) (2010) 317–329.
- [161] J. David, Algorithms for Analysis and Design of Robust Controllers, Ph.D. thesis, Kat. Univ. Leuven, 1994.
- [162] M. Fazel, H. Hindi, S.P. Boyd, Log-det heuristic for matrix rank minimization with applications to Hankel and Euclidean distance matrices, in: American Control Conference, vol. 3 2003, pp. 2156–2162.
- [163] K. Mohan, M. Fazel, Iterative reweighted algorithms for matrix rank minimization, *J. Mach. Learn. Res.* 13 (1) (2012) 3441–3473.
- [164] S.P. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University Press, London, 2004.
- [165] K.-C. Toh, M.J. Todd, R.H. Tütüncü, SDPT3—a MATLAB software package for semi-definite programming, version 1.3, *Optim. Methods Softw.* 11 (1–4) (1999) 545–581.
- [166] L. Vandenberghe, S. Boyd, Semidefinite programming, *SIAM Rev.* 38(1) (1996) 49–95.
- [167] S. Haghighatshoar, G. Caire, Massive MIMO channel subspace estimation from low-dimensional projections, *IEEE Trans. Signal Process.* 65 (2) (2017) 303–318.
- [168] C. Steffens, M. Pesavento, M.E. Pfetsch, A compact formulation for the  $\ell_{2,1}$  mixed-norm minimization problem, 2016, arXiv preprint arXiv:1606.07231.
- [169] D. Zachariah, M. Sundin, M. Jansson, S. Chatterjee, Alternating least-squares for low-rank matrix reconstruction, *IEEE Signal Process. Lett.* 19 (4) (2012) 231–234.
- [170] S. Bhojanapalli, A. Kyriolidis, S. Sanghavi, Dropping convexity for faster semi-definite optimization, 2015, arXiv preprint.
- [171] P. Jain, P. Netrapalli, S. Sanghavi, Low-rank matrix completion using alternating minimization, in: Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing, ACM, New York, NY, USA, 2013, pp. 665–674.
- [172] M.A. Davenport, J. Romberg, An overview of low-rank matrix recovery from incomplete observations, *IEEE J. Sel. Top. Signal Process.* 10 (4) (2016) 608–622.
- [173] S. Tu, R. Boczar, M. Soltanolkotabi, B. Recht, Low-rank solutions of linear matrix equations via procrustes flow, 2015, arXiv preprint arXiv:1507.03566.
- [174] D. Park, A. Kyriolidis, C. Caramanis, S. Sanghavi, Finding low-rank solutions to matrix problems, efficiently and provably, 2016, arXiv preprint arXiv:1606.03168.
- [175] M. Cho, J.-F. Cai, S. Liu, Y.C. Eldar, W. Xu, Fast alternating projected gradient descent algorithms for recovering spectrally sparse signals, in: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 4638–4642.
- [176] J.-F. Cai, T. Wang, K. Wei, Fast and provable algorithms for spectrally sparse signal reconstruction via low-rank Hankel matrix completion, 2016, arXiv preprint arXiv:1606.01567.
- [177] T. Yardibi, J. Li, P. Stoica, M. Xue, A.B. Baggerer, Source localization and sensing: a nonparametric iterative adaptive approach based on weighted least squares, *IEEE Trans. Aerosp. Electron. Syst.* 46 (1) (2010) 425–443.
- [178] K. Mahata, M.M. Hyder, Frequency estimation from arbitrary time samples, *IEEE Trans. Signal Process.* 64 (21) (2016) 5634–5643.
- [179] K. Mahata, M.M. Hyder, Grid-less TV minimization for DOA estimation, *Signal Process.* 132 (2017) 155–164.
- [180] B. Adcock, A.C. Hansen, Generalized sampling and infinite-dimensional compressed sensing, *Found. Comput. Math.* (2015) 1–61.

# Beamforming techniques using microphone arrays

# 12

**Rohith Mars\***, **Vaninirappuputhenpurayil Gopalan Reju\***, **Andy W.H. Khong\***,  
**Yusuke Hioka<sup>†</sup>**, **Kenta Niwa<sup>‡</sup>**

*Nanyang Technological University, Singapore*\* *University of Auckland, Auckland, New Zealand*<sup>†</sup>

*NTT Media Intelligence Laboratories, Tokyo, Japan*<sup>‡</sup>

## NOMENCLATURE

$(\cdot)^*$	complex conjugation operator
$w_m$	beamformer weight for the $m$ th microphone
$j$	$\sqrt{-1}$
$(\cdot)^H$	Hermitian operation for a matrix
$(\cdot)^T$	matrix transpose operator
$a$	steering vector of the microphone array
$w$	weight vector of the beamformer in time domain
$D$	directivity function matrix
$\Delta_m$	models the delay for the $m$ th microphone
$\epsilon$	regularization parameter
$\lambda$	signal wavelength
$\mathcal{F}$	frequency range of interest
$\mathcal{G}$	beamformer gain
$\mathcal{J}$	cost function
$\mu$	adaptation step size
$\Omega$	normalized frequency
$\omega$	signal angular frequency
$\omega_r$	reference frequency
$\phi$	power spectral density
$\phi_{Y_b}(k,l)$	PSD of the $b$ th beamformer output
$\tau_m$	signal propagation time delay between the first and the $m$ th microphone
$\Theta$	range of directions of interest
$\theta_i$	direction of arrival of the $i$ th source
$\Theta_m$	range of directions corresponding to mainlobe
$\Theta_s$	range of directions corresponding to sidelobe
$\tilde{w}$	response corresponding to constraint $c_j$
$\zeta$	sidelobe attenuation parameter

$c$	signal velocity
$D_{b, \theta}(k, l)$	directivity function of the beamformer $b$ for angle $\theta$
$h_{i, m}$	impulse response from the $i$ th source to the $m$ th microphone
$J$	number of filter taps at the output of one of the microphones
$P(\Omega)$	frequency response of FIR filter
$P(\omega, \theta)$	response of the beamformer
$t$	time index
$T_s$	signal sampling interval
$U(k, l)$	STFT coefficient of the Wiener filter output
$V(k, l)$	STFT coefficient of the noise signal
$v_m(t)$	noise of the $m$ th microphone
$X_m(k, l)$	STFT coefficient of the $m$ th microphone output
$x_m(t)$	the $m$ th microphone output at time instant $t$
$Y(k, l)$	STFT coefficient of the beamformer output
$y(t)$	beamformer output in time domain
$Z(k, l)$	STFT coefficient of the postfiltered beamformer output
$c_j$	column vectors of the constraint matrix $\mathbf{C}$
$H(k, l)$	vector of STFT coefficients of the impulse responses $H(k, l)$
$P(k)$	vector of beamformer responses $P_i^*(k)$ at frequency bin $k$
$R_{XX}$	covariance matrix of $\mathbf{X}(k, l)$ over $l$
$R_{xx}$	covariance matrix of $\mathbf{x}(t)$
$X(k, l)$	STFT coefficient vector of the microphone outputs
$x(t)$	signal vector at the output of the $M$ microphones at time instant $t$
$\Gamma_V$	noise pseudo coherence matrix
$*$	convolution operator
$\alpha$	forgetting factor
$\beta$	frequency-invariance parameter
$\sigma_x^2$	variance of random variable $x$
$\theta$	direction of arrival of the signal
$B$	total number of beamformers
$d$	spacing between adjacent microphones
$E[\cdot]$	expectation operator
$E_l[\cdot]$	expectation over $l$
$F(\omega, \theta)$	weighting function corresponding to wideband beamformer design
$m$	microphone index
$M$	number of microphones
$P_{\text{des}}(\omega, \theta)$	desired frequency response of the beamformer
$s(t)$	source signal at time instant $t$
$W_m(k, l)$	beamformer weight for the sample at index $(k, l)$ of the $m$ th microphone
$A(k)$	matrix of steering vectors
$\Sigma_V$	noise correlation matrix
$A_i(k)$	steering vector corresponding to direction $\theta_i$ for the $k$ th bin
$W(k, l)$	beamformer weight vector with elements $W_m(k, l)$
$I_M$	identity matrix of dimension $M \times M$

---

## 12.1 INTRODUCTION

Beamforming is an array signal processing technique for enhancing signals from one or more directions while suppressing noise and interferences from other directions using single or multiple sensor arrays. In audio beamforming microphones are being deployed as sensors. During the past few decades, beamformers have been widely used in radar, sonar, communication, speech processing, medical imaging, surveillance, and other areas [1–3]. Applications of the technique include source localization, where the direction-of-arrival (DOA) of signal sources is estimated based on the energy extracted from each direction [4–7], and signal enhancement, in which the surrounding noise and interferences are suppressed and signals from the direction of interest are preserved [1, 6].

Spatial filtering is a method similar to temporal filtering [6]. In temporal filtering, when the frequency of the desired signal and that of the interfering signal lie in the same frequency band, removal of the interference is challenging especially in the presence of a large number of such interfering signals. However, when the desired signal and the interfering signals originate from different directions, the interfering signals can be suppressed from the desired signals through spatial filtering. In temporal filtering, the signals are sampled in temporal aperture whereas in spatial filtering the signals will be sampled by spatially located multiple sensors. The spatially sampled signals will then be linearly combined to obtain the spatially filtered output (beamformer output). The spatial discrimination capability of the beamformer can be altered by changing the spatial configuration of the sensors. These sensor configurations can be regular or nonregular. In a regular configuration, the sensors will be placed with uniform or nonuniform spacing whereas in nonregular configuration, they will be placed at random positions. Different types of array configurations such as linear and planar (rectangular/circular) arrays, and three-dimensional arrays can be employed depending on scenario and usage [4]. In this chapter, our focus is on beamforming algorithms which use microphone arrays of regular spacing for audio applications and hence wideband beamforming.

This chapter is organized as follows: In [Section 12.2](#), we start with the mathematical formulation of beamforming technique leading to delay-and-sum beamforming (DSB) and its wideband version namely filter-and-sum beamforming. We describe, in [Section 12.3](#), the most widely used adaptive and nonadaptive beamforming techniques including the superdirective (SD) beamformer [8–10], linearly constraint minimum variance (LCMV) beamformer [11, 12], and minimum variance distortionless response (MVDR) beamformer [13–15]. The desired signal obtained at the beamformer output is generally distorted due to reasons such as errors in estimation of signal statistics, modeling errors and environmental noise. Hence, a postfiltering is generally carried out on the beamformer output to improve the quality of the desired signal. Therefore, in [Section 12.4](#) we describe a power spectrum density (PSD) estimation based postfiltering technique. Finally [Section 12.5](#) concludes this chapter.

## 12.2 PROBLEM FORMULATION

In this section, we present the terminologies and describe the fundamental concepts of beamforming. For ease of understanding, we first describe the concepts in narrowband beamforming where we deal with signals of a single frequency and describe the determination of beamformer weights from a finite impulse response (FIR) filter design perspective. Using the concepts introduced for narrowband beamforming, we then extend the discussion for wideband signals such as speech.

### 12.2.1 NARROWBAND BEAMFORMING

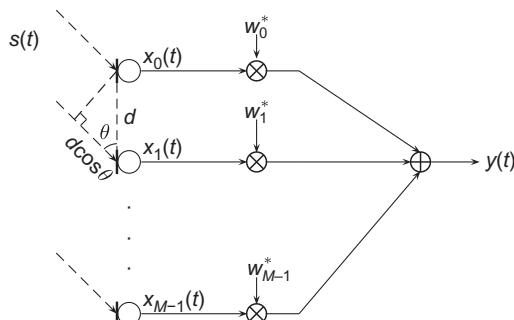
For ease of explanation, we assume an array of  $M$  linearly arranged microphones with a uniform spacing of  $d$  between two adjacent microphones as shown in Fig. 12.1. We also assume a signal  $s(t)$  arriving at the microphones from an angle  $\theta$ . The source is positioned at a far distance (compared to its wavelength) such that the propagating waves can be approximated as planar [4]. Due to the far-field assumption and uniform spacing across the microphones, the delay between signals received at the first microphone and the  $m$ th microphone is given by [4]

$$\tau_m(\theta) = \frac{md\cos\theta}{c}, \quad m = 0, 1, \dots, M-1, \quad (12.1)$$

where  $c$  is the velocity of sound propagation. Hence the signal corresponding to the  $m$ th microphone output can be written as

$$\begin{aligned} x_m(t) &= s(t - \tau_m(\theta)) \\ &= \Delta_m s(t), \end{aligned} \quad (12.2)$$

where  $\Delta_m$  is a signal dependent model (delay) which when applied on the reference signal corresponds to the  $m$ th microphone output. Note that in the following



**FIG. 12.1**

The delay-and-sum beamformer.

discussion, we will omit  $\theta$  in  $\tau_m(\theta)$  for notational brevity unless it is necessary. Considering all microphones, the  $M \times 1$  vector of received signals is given by

$$\mathbf{x}(t) = \mathbf{a}s(t), \quad (12.3)$$

where  $\mathbf{x}(t) = [x_0(t), x_1(t), \dots, x_{M-1}(t)]^T$  and  $\mathbf{a} = [1, \Delta_1, \dots, \Delta_{M-1}]^T$  is the delay vector governed by actual wave propagation such that premultiplying  $s(t)$  by  $\mathbf{a}$  will delay the signal  $s(t)$ . The objective of beamforming is to estimate  $s(t)$  from the observed (received) signals  $\mathbf{x}(t)$ .

To design the beamformer, we first note, with reference to Eq. (12.3), that  $s(t)$  can be expressed in terms of  $\mathbf{x}(t)$  given by [16]

$$s(t) = (\mathbf{a}^H \mathbf{a})^{-1} \mathbf{a}^H \mathbf{x}(t). \quad (12.4)$$

For illustrative purpose, we assume that the source signal is complex and is given by  $s(t) = e^{j\omega t}$  with an angular frequency of  $\omega$ . The output of the  $m$ th microphone can therefore be written as

$$\begin{aligned} x_m(t) &= e^{j\omega(t-\tau_m)} \\ &= e^{-j\omega\tau_m} e^{j\omega t} \\ &= \Delta_m s(t). \end{aligned} \quad (12.5)$$

Comparing Eqs. (12.5) and (12.3), and noting that  $\tau_0 = 0$ , the delay vector is given by  $\mathbf{a} = [1, e^{-j\omega\tau_1}, \dots, e^{-j\omega\tau_{M-1}}]^T$ . With reference to Eq. (12.4), and letting  $(\mathbf{a}^H \mathbf{a})^{-1} = 1/M$  be the scaling factor, the relationship between the source and received signal is given by

$$\begin{aligned} s(t) &= \frac{1}{M} \mathbf{a}^H \mathbf{x}(t) \\ &= \frac{1}{M} [1, e^{j\omega\tau_1}, \dots, e^{j\omega\tau_{M-1}}] \begin{bmatrix} x_0(t) \\ x_1(t) \\ \vdots \\ x_{M-1}(t) \end{bmatrix}. \end{aligned} \quad (12.6)$$

The design of a typical delay-and-sum beamformer involves the use of weights  $\mathbf{w} = [w_0, w_1, \dots, w_{M-1}]^T$  such that [6]

$$y(t) = \mathbf{w}^H \mathbf{x}(t). \quad (12.7)$$

Comparing Eqs. (12.6) and (12.7), and to achieve  $y(t) \approx s(t)$ , the beamformer weight vector is given by

$$\mathbf{w}^H = [w_0^*, w_1^*, \dots, w_{M-1}^*], \quad (12.8)$$

$$\text{where } w_m^* = \frac{e^{j\omega\tau_m}}{M}.$$

From the above discussion it is clear that if one multiplies the microphone outputs with appropriate complex scalars and sums them up, one can, in theory, recover  $s(t)$  emanating from direction  $\theta$ . This is the operating principle of the delay-and-sum beamformer where signals from the desired direction are constructively added to achieve (desired) signal enhancement while signals from all other directions are

destructively added resulting in interference suppression. This implies that the response of a beamformer with fixed weights (i.e., the beamformer designed to enhance signals from only one direction and suppress signals from all other directions) will vary across directions. Similarly, since the weights of the beamformer are designed to enhance signals of a particular angular frequency  $\omega$ , the response will vary across frequencies for the same direction.

To characterize the beamformer, it is important to determine the beamformer's response for various frequencies as a function of  $\theta$ . Substituting Eq. (12.3) into Eq. (12.7), we have

$$\begin{aligned} y(t) &= \mathbf{w}^H \mathbf{a}(\omega, \theta) s(t) \\ &= P(\omega, \theta) s(t), \end{aligned} \quad (12.9)$$

where

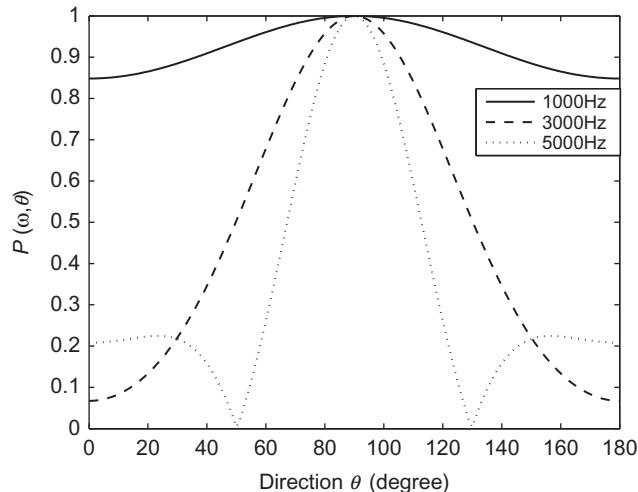
$$\begin{aligned} P(\omega, \theta) &= \mathbf{w}^H \mathbf{a}(\omega, \theta) \\ &= \sum_{m=0}^{M-1} w_m^* e^{-j\omega\tau_m} \end{aligned} \quad (12.10)$$

is defined as the beamformer's response, and

$$\mathbf{a}(\omega, \theta) = [1, e^{-j\omega\tau_1}, \dots, e^{-j\omega\tau_{M-1}}]^T \quad (12.11)$$

is generally referred to as the steering vector of the array [6].

**Fig. 12.2** shows the response of a 10-microphone delay-and-sum beamformer for signals of three different frequencies. The spacing between the microphones has been fixed at 1/4 of the wavelength of the highest frequency component. We note that the



**FIG. 12.2**

Response of the delay-and-sum beamformer.

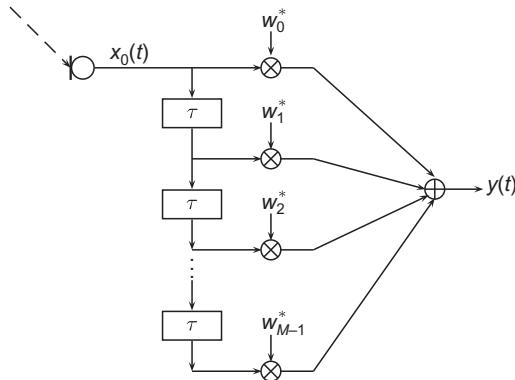


FIG. 12.3

Delay-and-sum beamformer with microphones replaced by delay elements.

higher frequency components can achieve better spatial resolution compared to lower frequencies. We also note, from Eq. (12.10) and Fig. 12.4 that the beamformer response is a function of  $M$  and  $d$  such that a high resolution can be achieved with a high  $M$  and large  $d$ .

**Spatial aliasing:** Since signals received at the microphones are delayed versions of the signal at the first microphone, we can re-draw Fig. 12.1 by replacing the last  $M - 1$  microphones with delay elements as shown in Fig. 12.3. This structure bears resemblance to an FIR filter. In an FIR filter, the tap delay corresponds to the sampling time or spacing between two adjacent samples in seconds whereas in beamforming, it corresponds to the signal propagation delay between adjacent microphones. This implies that the signal is spatially sampled in a manner that is similar to the temporal sampling of signals. In temporal sampling, if the sampling frequency is less than twice the highest frequency of the signal, the resultant sample sequence will be interpreted as that of some other frequency component. This is referred to as signal aliasing in the temporal sampling process [17]. Similarly for the spatial sampling case, the signal must be sampled with  $d \leq \lambda/2$ , to avoid spatial aliasing. Here,  $\lambda$  is the wavelength corresponding to the highest frequency of the signal.

If  $d > \lambda/2$ , spatial aliasing will occur and hence sources at different directions will have the same response/steering vector. This implies that, for the same angular frequency  $\omega$ ,

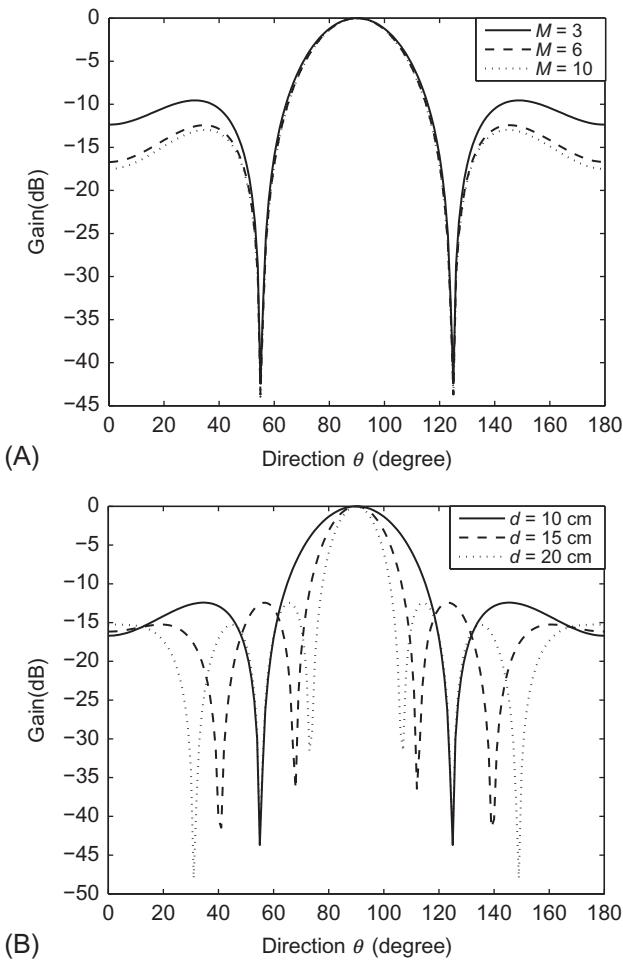
$$\mathbf{a}(\omega, \theta_1) = \mathbf{a}(\omega, \theta_2), \quad (12.12)$$

or

$$e^{-j\omega\tau_m(\theta_1)} = e^{-j\omega\tau_m(\theta_2)}, \quad (12.13)$$

where  $\tau_m(\theta_i)$  is the delay corresponding to a given direction  $\theta_i$ . Substituting  $\tau_m = m\tau_1$ ,  $\tau_1 = \frac{d\cos\theta}{c}$ , and  $f/c = 1/\lambda$ , we have

$$e^{-jm(2\pi d\cos\theta_1)/\lambda} = e^{-jm(2\pi d\cos\theta_2)/\lambda}. \quad (12.14)$$

**FIG. 12.4**

Beampatterns at  $f = 1 \text{ kHz}$  showing the (A) effect of microphone number  $M$  with  $d = 60/M \text{ cm}$  and (B) intermicrophone spacing  $d$ , keeping  $M = 6$ .

To avoid aliasing, the magnitudes of both  $(2\pi d \cos \theta_1)/\lambda$  and  $(2\pi d \cos \theta_2)/\lambda$  must be less than  $\pi$ . Since  $|\cos \theta| \leq 1$ , the above condition therefore requires that  $d \leq \lambda/2$ . As such we note that, to avoid spatial aliasing, the spacing between the microphones must be less than half the wavelength (corresponding to the highest frequency) of the source signal.

From Fig. 12.3 one may note the similarity between the delay-and-sum beamformer and the FIR filter. More specifically and for ease of explanation, let us assume that  $d = \lambda/2$  so that the beamformer response will be simplified to

$$P(\omega, \theta) = \sum_{m=0}^{M-1} e^{-jm\pi \cos \theta} w_m^*. \quad (12.15)$$

For an FIR filter with the same set of coefficients,  $w_m^*$ ,  $m = 0, 1, \dots, M - 1$ , the filter response will be [18]

$$P(\Omega) = \sum_{m=0}^{M-1} e^{-jm\Omega} w_m^*, \quad (12.16)$$

where  $-\pi \leq \Omega \leq \pi$  is the normalized frequency. It can be seen, from Eq. (12.15), that if the direction of the source  $\theta$  changes from  $\pi$  to 0, the term  $\pi \cos \theta$  will also change from  $-\pi$  to  $\pi$  and hence Eqs. (12.15), (12.16) are equivalent. Therefore, any FIR design methodologies that are used for the estimation of the filter coefficients can be employed for the estimation of delay-and-sum beamformer weights such that the desired direction corresponds to the pass-band of the FIR filter and the stop-band region corresponds to the undesired directions. For the case where  $d < \eta\lambda/2$  for  $0 < \eta < 1$ , Eq. (12.15) will be modified to

$$P(\omega, \theta) = \sum_{m=0}^{M-1} e^{-jm\eta\pi \cos \theta} w_m^*. \quad (12.17)$$

The beamformer coefficients can therefore be designed in the same way as before except that the response of the FIR filter for the region  $\Omega \in [-\pi, -\eta\pi]$  and  $\Omega \in [\eta\pi, \pi]$  will be arbitrary [3].

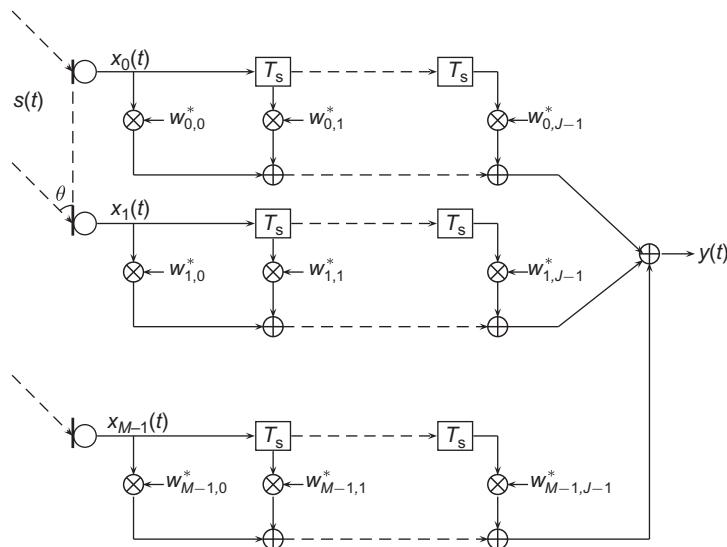
### 12.2.2 WIDEBAND BEAMFORMING

From Eq. (12.6) and Fig. 12.2, we note that for a fixed  $d$  the beamformer weights are frequency dependent. However, signals such as speech are wideband signals, i.e., they contain large number of frequency components. Hence, to achieve good performance, different weights are required across various frequencies. An immediate extension of narrowband beamforming for a wideband signal is achieved by first splitting the components of the wideband signal into different frequencies and subsequently using appropriate weights for each frequency component. One of the easiest methods to achieve this is through the use of the Fourier transform where the signal is first converted to the frequency domain and then multiplying the components with the corresponding weights.

To illustrate the above, we let  $\mathbf{X}(k, l) = [X_0(k, l), X_1(k, l), \dots, X_{M-1}(k, l)]^T$  be the short-time Fourier transform (STFT) of the microphone outputs. The output of the beamformer in the STFT domain can be written as

$$Y(k, l) = \mathbf{W}^H(k) \mathbf{X}(k, l), \quad (12.18)$$

where  $\mathbf{W}(k)$  is the weight vector,  $\omega_k = 2\pi \frac{(k-1)}{K} f_s$ ,  $k$  is the STFT frequency-bin index,  $K$  is the DFT length,  $f_s$  is the sampling frequency, and  $l$  is the time-frame index.

**FIG. 12.5**

The filter-and-sum beamformer.

The above can be implemented in the time domain by replacing the single weights in the delay-and-sum beamformer by an FIR filter. This results in the well-known filter-and-sum beamformer, as shown in Fig. 12.5 [6]. The reason why such a filter can be used for wideband beamforming is due to the temporal filtering process of the FIR filter which imposes a frequency-dependent response on the received wideband signal. This implies that the response of the filter varies with frequency. Therefore, by appropriately selecting the FIR filter coefficients which meets the desired response for all the frequency components present in the wideband signal, it is possible to spatially filter the received wideband signal.

To describe the filter-and-sum beamformer, we note that its output is given by

$$y(t) = \sum_{m=0}^{M-1} \sum_{j=0}^{J-1} x_m(t - jT_s) w_{m,j}^*, \quad (12.19)$$

where  $J$  is the filter length for each microphone,  $T_s$  is the delay between adjacent taps of the FIR filter and  $w_{m,j}$  is the weight coefficient of the  $m$ th microphone at  $j$ th tap. This delay corresponds to the temporal sampling interval. Eq. (12.19) may be written in vector form as

$$y(t) = \mathbf{w}^H \mathbf{x}(t), \quad (12.20)$$

where the  $MJ \times 1$  weight vector  $\mathbf{w}$  is given by

$$\mathbf{w} = [w_{0,0}, \dots, w_{M-1,0}, \dots, w_{0,J-1}, \dots, w_{M-1,J-1}]^T \quad (12.21)$$

and the data vector  $\mathbf{x}(t)$  is given by

$$\mathbf{x}(t) = [\mathbf{x}_0(t), \mathbf{x}_1(t - T_s), \dots, \mathbf{x}_{J-1}(t - (J-1)T_s)]^T, \quad (12.22)$$

with

$$\mathbf{x}_j(t - jT_s) = [x_0(t - jT_s), x_1(t - jT_s), \dots, x_{M-1}(t - jT_s)]^T, \quad j = 0, 1, \dots, J-1, \quad (12.23)$$

as shown in Fig. 12.5 [11].

Now for an arbitrarily defined complex plane wave  $e^{j\omega t}$  impinging on the microphones and assuming that the first microphone is the reference microphone such that  $x_0(t) = e^{j\omega t}$ , we have

$$x_m(t - jT_s) = e^{j\omega(t - (\tau_m + jT_s))}. \quad (12.24)$$

With reference to Fig. 12.5, the array output can then be written as

$$\begin{aligned} y(t) &= e^{j\omega t} \sum_{m=0}^{M-1} \sum_{j=0}^{J-1} e^{-j\omega(\tau_m + jT_s)} w_{m,j}^* \\ &= e^{j\omega t} P(\omega, \theta), \end{aligned} \quad (12.25)$$

where, similar to the narrowband beamformer,

$$P(\theta, \omega) = \mathbf{w}^H \mathbf{a}(\omega, \theta) \quad (12.26)$$

is the beamformer response and

$$\mathbf{a}(\omega, \theta) = \left[ e^{-j\omega\tau_0}, \dots, e^{-j\omega\tau_{M-1}}, e^{-j\omega(\tau_0 + T_s)}, \dots, e^{-j\omega(\tau_{M-1} + T_s)}, \dots, e^{-j\omega(\tau_0 + (J-1)T_s)}, \dots, e^{-j\omega(\tau_{M-1} + (J-1)T_s)} \right]^T \quad (12.27)$$

is the steering vector of the wideband beamformer.

The weights of the filter-and-sum beamformer can be estimated by techniques such as convex optimization [19], the method of least squares (LS) [20, 21], and the eigenfilter method [22]. For ease of explanation, we describe a wideband design method using least squares. In this technique, the weighted sum of the squared error between a desired response and beamformer response is minimized over a frequency range of interest and direction. The general form of a weighted least squares cost function for the design of a wideband beamformer is given by [23]

$$\mathcal{J}_{\text{LS}} = \int_{\mathcal{F}} \int_{\Theta} F(\omega, \theta) |P(\omega, \theta) - P_{\text{des}}(\omega, \theta)|^2 d\omega d\theta, \quad (12.28)$$

where  $|\cdot|^2$  denotes the squared 2-norm,  $\mathcal{F}$  and  $\Theta$  denote the frequency range of interest and directions, respectively. The variable  $F(\omega, \theta)$  is a positive real-valued weighting function corresponding to the mainlobe and sidelobes of the beampattern while  $P(\omega, \theta)$  and  $P_{\text{des}}(\omega, \theta)$  define the actual and desired directional responses, respectively. By selecting a uniform grid of frequencies and direction angles, we assume  $P_{\text{des}}(\omega, \theta) = 1$  and  $F(\omega, \theta) = 1$  for the mainlobe. For the sidelobes, we assume  $P_{\text{des}}(\omega, \theta) = 0$  and  $F(\omega, \theta) = \zeta$ ,  $0 < \zeta < 1$ . The cost function in Eq. (12.28) can then be rewritten, using Eq. (12.26), as

$$\mathcal{J}_{\text{LS}} = \sum_{\omega \in \mathcal{F}} \sum_{\theta \in \Theta_m} |\mathbf{w}^H \mathbf{a}(\omega, \theta) - 1|^2 + \zeta \sum_{\omega \in \mathcal{F}} \sum_{\theta \in \Theta_s} |\mathbf{w}^H \mathbf{a}(\omega, \theta)|^2, \quad (12.29)$$

where  $\Theta_m$  and  $\Theta_s$  denote the range of directions corresponding to mainlobe and sidelobes, respectively. Eq. (12.29) is solved using the Lagrange multiplier method to obtain the beamformer weights as

$$\mathbf{w}_{LS} = \mathbf{Q}_{LS}^{-1} \mathbf{a}_{LS}, \quad (12.30)$$

where

$$\mathbf{Q}_{LS} = \sum_{\omega \in \mathcal{F}} \sum_{\theta \in \Theta_m} \mathbf{a}(\omega, \theta) \mathbf{a}^H(\omega, \theta) + \zeta \sum_{\omega \in \mathcal{F}} \sum_{\theta \in \Theta_s} \mathbf{a}(\omega, \theta) \mathbf{a}^H(\omega, \theta) \quad (12.31)$$

and

$$\mathbf{a}_{LS} = \sum_{\omega \in \mathcal{F}} \sum_{\theta \in \Theta_m} \mathbf{a}(\omega, \theta). \quad (12.32)$$

It can be seen that by using the formulation in Eq. (12.29), we obtain a frequency-dependent beamformer. However, to achieve a frequency-invariant beamformer (FIB), with constant beamwidth across all the frequencies, additional constraints have to be introduced. To achieve such a frequency-invariant property, the “response variation” criterion has been proposed [24]. This criterion defines the Euclidean distance between the response at a reference frequency  $\omega_r$  and that of all frequencies over the range of directions  $\Theta_{FI}$ . The cost function to be minimized to achieve this criterion is given by

$$\mathcal{J}_{rv} = \sum_{\omega \in \mathcal{F}} \sum_{\theta \in \Theta_{FI}} |\mathbf{w}^H \mathbf{a}(\omega, \theta) - \mathbf{w}^H \mathbf{a}(\omega_r, \theta)|^2. \quad (12.33)$$

This cost function along with Eq. (12.29) offers a tradeoff between sidelobe attenuation and the frequency-invariance property. This tradeoff can be varied through the use of a control parameter  $\beta$  such that the resultant frequency-invariant cost function is given by

$$\mathcal{J}_{FIB} = \mathcal{J}_{LS} + \beta \mathcal{J}_{rv}. \quad (12.34)$$

It is useful to note that the frequency-invariant property can be obtained either in the mainlobe direction or the full range of directions. If it is considered over the entire range of directions, i.e., including the sidelobe region, the spectrum energy of the beamformer needs to be minimized only at the reference frequency  $\omega_r$  instead of over the entire frequency range.

Substituting  $\mathcal{J}_{LS}$  from Eq. (12.29) into Eq. (12.34), the modified FIB cost function can be expressed as

$$\mathcal{J}_{FIB} = \sum_{\theta \in \Theta_m} |\mathbf{w}^H \mathbf{a}(\omega_r, \theta) - 1|^2 + \zeta \sum_{\theta \in \Theta_s} |\mathbf{w}^H \mathbf{a}(\omega_r, \theta)|^2 + \beta \mathcal{J}_{rv},$$

where  $\zeta$  is the sidelobe attenuation parameter. The above cost function can also be solved using the Lagrange multiplier method to obtain the optimum solution as [20]

$$\mathbf{w} = \mathbf{Q}_{FIB}^{-1} \mathbf{a}_{FIB}, \quad (12.35)$$

where

$$\begin{aligned} \mathbf{Q}_{FIB} &= \sum_{\theta \in \Theta_m} \mathbf{a}(\omega_r, \theta) \mathbf{a}^H(\omega_r, \theta) + \zeta \sum_{\theta \in \Theta_s} \mathbf{a}(\omega_r, \theta) \mathbf{a}^H(\omega_r, \theta) \\ &\quad + \beta \sum_{\omega \in \mathcal{F}} \sum_{\theta \in \Theta} (\mathbf{a}(\omega, \theta) - \mathbf{a}(\omega_r, \theta)) (\mathbf{a}(\omega, \theta) - \mathbf{a}(\omega_r, \theta))^H \end{aligned} \quad (12.36)$$

and the FIB steering vector is given by

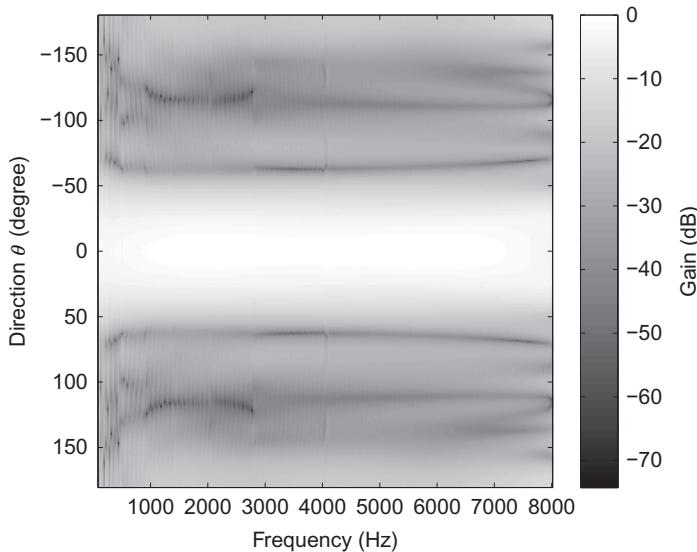
$$\mathbf{a}_{\text{FIB}} = \sum_{\theta \in \Theta_m} \mathbf{a}(\omega_r, \theta). \quad (12.37)$$

**Fig. 12.6** shows the beampattern using the LS technique [20] for a linear array of  $M = 10$  microphones, each having  $J = 200$  tap filters with  $d = 2$  cm and  $\theta = 0^\circ$ . The sidelobe attenuation parameter  $\zeta$  and the frequency-invariance tradeoff parameter  $\beta$  is set at 0.01 and 0.95, respectively. It can be seen that frequency-invariance is achieved across all the frequencies.

### 12.3 BASIC APPROACHES IN WIDEBAND BEAMFORMING

While formulating the beamformer, we have assumed, in previous sections, that the environment is ideal, i.e., nonreverberant and noise-free environment. For a practical scenario, we need to formulate the beamforming problem by taking room reverberation, interferences and environmental noise into account. In this case, the received signal is given by

$$x_m(t) = \sum_{i=0}^I h_{i,m}(t) * s_i(t) + v_m(t), \quad (12.38)$$



**FIG. 12.6**

The beampattern of frequency-invariant beamformer using the LS technique for a linear array of  $M = 10$  microphones with  $d = 2$  cm and  $\theta = 0^\circ$ .

where  $*$  denotes the convolution operator. The variable  $s_i(t)$  denotes the desired source for  $i = 0$  and interferers for  $i \neq 0$ ,  $v_m(t)$  denotes additive noise of the  $m$ th microphone, and  $h_{i,m}(t)$  is the impulse response (IR) from the  $i$ th source to the  $m$ th microphone. To facilitate modeling wideband signal such as human speech, it is common to process these signals in the time-frequency (TF) domain. Hence, taking the STFT, the above equations can be expressed in the TF domain as

$$\begin{aligned}\mathbf{X}(k,l) &= \sum_{i=0}^I \mathbf{H}_i(k,l) S_i(k,l) + \mathbf{V}(k,l) \\ &= [X_0(k,l), X_1(k,l), \dots, X_{M-1}(k,l)]^T,\end{aligned}\quad (12.39)$$

where

$$\mathbf{V}(k,l) = [V_0(k,l), V_1(k,l), \dots, V_{M-1}(k,l)]^T, \quad (12.40)$$

$$\mathbf{H}_i(k,l) = [H_{i,0}(k,l), H_{i,1}(k,l), \dots, H_{i,M-1}(k,l)]^T, \quad (12.41)$$

and  $S_i(k, l)$  is the STFT coefficient of the  $i$ th source signal corresponding to an angle  $\theta_i$ . The  $M \times 1$  vectors  $\mathbf{X}(k, l)$  and  $\mathbf{V}(k, l)$  are the STFT coefficients of the microphone outputs and the additive noise, respectively, while  $H_{i,m}(k, l)$  is the STFT coefficient of the impulse response from  $i$ th source to the  $m$ th microphone. If the direction of the desired signal does not change with time and the incoming signals are assumed to be plane waves, the vector corresponding to the impulse response  $\mathbf{H}_i(k, l)$  can be replaced by an  $M \times 1$  steering vector  $\mathbf{A}_i(k) = [A_{i,0}(k), A_{i,1}(k), \dots, A_{i,M-1}(k)]^T$ , provided that the environment is anechoic. While reverberation has to be taken into account, we assume that the environment is anechoic temporarily for clarity. The components of  $\mathbf{A}_i(k)$  are then given as

$$A_{i,m}(k) = \exp \left[ j2\pi f_k \frac{md}{c} \cos \theta_i \right]. \quad (12.42)$$

Defining  $S_0(k, l)$  as the STFT coefficient of the desired signal, the beamformer algorithm aims to estimate  $Y(k, l) \approx S_0(k, l)$  by applying weight vector  $\mathbf{W}(k, l)$  on the received signal as

$$Y(k, l) = \mathbf{W}^H(k, l) \mathbf{X}(k, l), \quad (12.43)$$

where

$$\mathbf{W}(k, l) = [W_0(k, l), W_1(k, l), \dots, W_{M-1}(k, l)]. \quad (12.44)$$

It may be noted that the weight vector  $\mathbf{W}(k, l)$  in Eq. (12.43) is a time-dependent variable and hence  $\mathbf{W}(k, l)$  is used instead of  $\mathbf{W}(k)$ . However, depending on whether the beamforming algorithm is data dependent or independent, these weights may or may not vary with time. For instance, beamformers which are designed based on only the source direction and array geometry will have fixed weights while beamformers which utilize the signal statistics for weight estimation will have time-varying weights that depend on the variation in statistics of the signals.

An important parameter used for the design of a beamformer is the beamformer gain [4]. From Eq. (12.43) (assuming an anechoic environment) we have

$$Y(k, l) = \mathbf{W}^H(k, l) \{ \mathbf{A}_0(k) S_0(k, l) + \mathbf{V}(k, l) \}. \quad (12.45)$$

Using the first microphone as the reference sensor, we can define the input signal-to-noise ratio (SNR) as

$$\text{iSNR}(k, l) = \frac{\sigma_{S_0}^2(k, l)}{\sigma_{V_0}^2(k, l)}, \quad (12.46)$$

where  $\sigma_{S_0}^2(k, l) = E[|S_0(k, l)|^2]$  and  $\sigma_{V_0}^2(k, l) = E[|V_0(k, l)|^2]$  are the variances of  $S_0(k, l)$  and  $V_0(k, l)$ , respectively. The output SNR is then defined as

$$\begin{aligned} \text{oSNR}[\mathbf{W}(k, l)] &= \frac{(\mathbf{W}^H(k, l) \mathbf{A}_0(k) S_0(k, l))^2}{(\mathbf{W}^H(k, l) \mathbf{V}(k, l))^2} \\ &= \sigma_{S_0}^2(k, l) \times \frac{|\mathbf{W}^H(k, l) \mathbf{A}_0(k)|^2}{\mathbf{W}^H(k, l) \Sigma_V(k, l) \mathbf{W}(k, l)} \\ &= \frac{\sigma_{S_0}^2(k, l)}{\sigma_{V_0}^2(k, l)} \times \frac{|\mathbf{W}^H(k, l) \mathbf{A}_0(k)|^2}{\mathbf{W}^H(k, l) \Gamma_V(k, l) \mathbf{W}(k, l)}, \end{aligned} \quad (12.47)$$

where  $\Sigma_V(k, l)$  and  $\Gamma_V(k, l) = \frac{\Sigma_V(k, l)}{\sigma_{V_0}^2(k, l)}$  denote the correlation and pseudo-coherence matrix of  $\mathbf{V}(k, l)$ , respectively. From Eqs. (12.46), (12.47), the gain of the beamformer in terms of SNR can be obtained as

$$\begin{aligned} \mathcal{G}[\mathbf{W}(k, l)] &= \frac{\text{oSNR}[\mathbf{W}(k, l)]}{\text{iSNR}(k, l)} \\ &= \frac{|\mathbf{W}^H(k, l) \mathbf{A}_0(k)|^2}{\mathbf{W}^H(k, l) \Gamma_V(k, l) \mathbf{W}(k, l)}. \end{aligned} \quad (12.48)$$

For any beamformer design, one of the main objectives is to estimate  $\mathbf{W}(k, l)$  that maximizes the beamformer gain. As an illustrative example to see how wideband beamforming can be realized using delay-and-sum beamformer, consider the case where the noise  $v_m(t)$ ,  $\forall m$  is white. We can model  $\Gamma_V(k, l) = \mathbf{I}_M$ , where  $\mathbf{I}_M$  is the  $M \times M$  identity matrix. Eq. (12.48) will then be simplified to

$$\mathcal{G}[\mathbf{W}(k)] = \frac{|\mathbf{W}^H(k) \mathbf{A}_0(k)|^2}{\mathbf{W}^H(k) \mathbf{W}(k)}. \quad (12.49)$$

Since the noise is considered to be white, the gain in Eq. (12.49) is defined as white noise gain (WNG). The WNG denotes the gain of the beamformer for the desired signal from the desired direction relative to the white noise amplification. When  $\mathcal{G}[\mathbf{W}(k)]$  is less than one, the white noise is amplified at the beamformer output. Note

that in Eq. (12.49), the index  $l$  is omitted as the delay-and-sum beamformer is independent of data. Maximizing Eq. (12.49) with the constraint for distortionless response across all frequencies, i.e.,  $\mathbf{W}^H(k)\mathbf{A}_0(k) = 1$ , and solving using Lagrange multiplier method, the weights for the  $k$ th frequency bin can be obtained as,

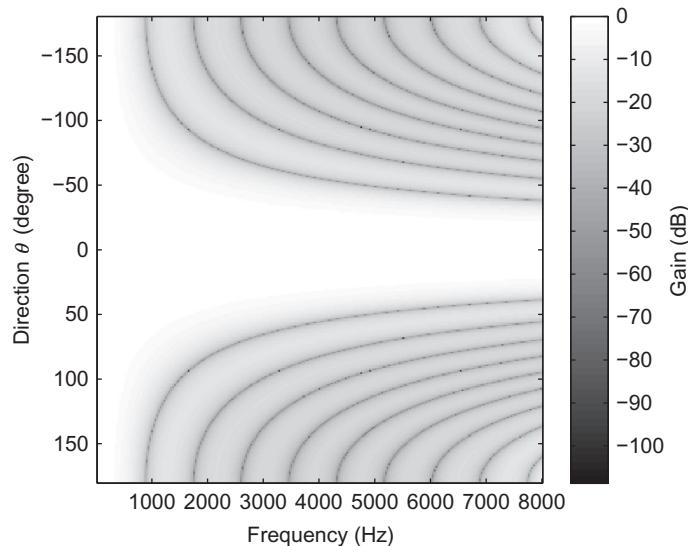
$$\mathbf{W}(k) = \frac{\mathbf{A}_0(k)}{M}. \quad (12.50)$$

This implies that the wideband beamforming can be realized using  $K$  delay-and-sum beamformers, one for each frequency bin. It is interesting to note that the weights obtained in Eq. (12.50) are same as those in Eq. (12.6), which was formulated in time domain. Fig. 12.7 shows a typical beampattern of the above beamformer using a linear array. Here we used  $M = 10$ ,  $d = 2$  cm and  $\theta_0 = 0^\circ$ .

### 12.3.1 SUPERDIRECTIVE BEAMFORMER

While deriving the weights for the above beamformer, we temporarily assumed that the environment is anechoic with additive white noise at the microphone outputs. However, in practical application environments such as rooms, the effect of reverberation cannot be neglected and the noise need not only be white. In a reverberant environment at a distance from the source exceeding the critical distance, the noise field is generally assumed to be diffuse. Considering such a diffuse noise field, the  $M \times M$  matrix,  $\Gamma_V(k)$  can be re-modeled as [25]

$$[\Gamma_V(k)]_{ij} = \text{sinc}[2\pi f_k(j-i)d/c], \quad (12.51)$$



**FIG. 12.7**

The beampattern of the delay-and-sum-based wideband beamformer using a linear array of  $M = 10$  microphones with  $d = 2$  cm and  $\theta_0 = 0^\circ$ .

which is data-independent and hence the index  $l$  is omitted. Substituting Eq. (12.51) in Eq. (12.48), the beamformer gain will be

$$\mathcal{G}[\mathbf{W}(k)] = \frac{|\mathbf{W}^H(k)\mathbf{A}_0(k)|^2}{\mathbf{W}^H(k)\Gamma_V(k)\mathbf{W}(k)}. \quad (12.52)$$

Similar to the weight estimation method used for the delay-and-sum-based wideband beamformer in the previous section,  $\mathbf{W}(k)$  can be estimated by maximizing Eq. (12.52) with the constraint  $\mathbf{W}^H(k)\mathbf{A}_0(k) = 1$  for distortionless response in the desired direction. This can be solved using the Lagrange multiplier method to obtain [9]

$$\mathbf{W}(k) = \frac{\Gamma_V^{-1}(k)\mathbf{A}_0(k)}{\mathbf{A}_0^H(k)\Gamma_V^{-1}(k)\mathbf{A}_0(k)}. \quad (12.53)$$

In Eq. (12.52), since we have considered the diffused noise field and maximize the gain in the direction of the desired signal, the beamformer gain in this case is generally known as the directivity factor (DF) and the beamformer obtained is known as the superdirective (SD) beamformer.

It can be seen that by substituting Eq. (12.53) in Eq. (12.49), the corresponding WNG will be less than one, which implies that the white noise will be amplified by the SD beamformer which limits its application in practice. To address this issue, various approaches have been proposed. In [9, 26] the DF is maximized subject to a constraint in Eq. (12.49). The resulting solution is then given by

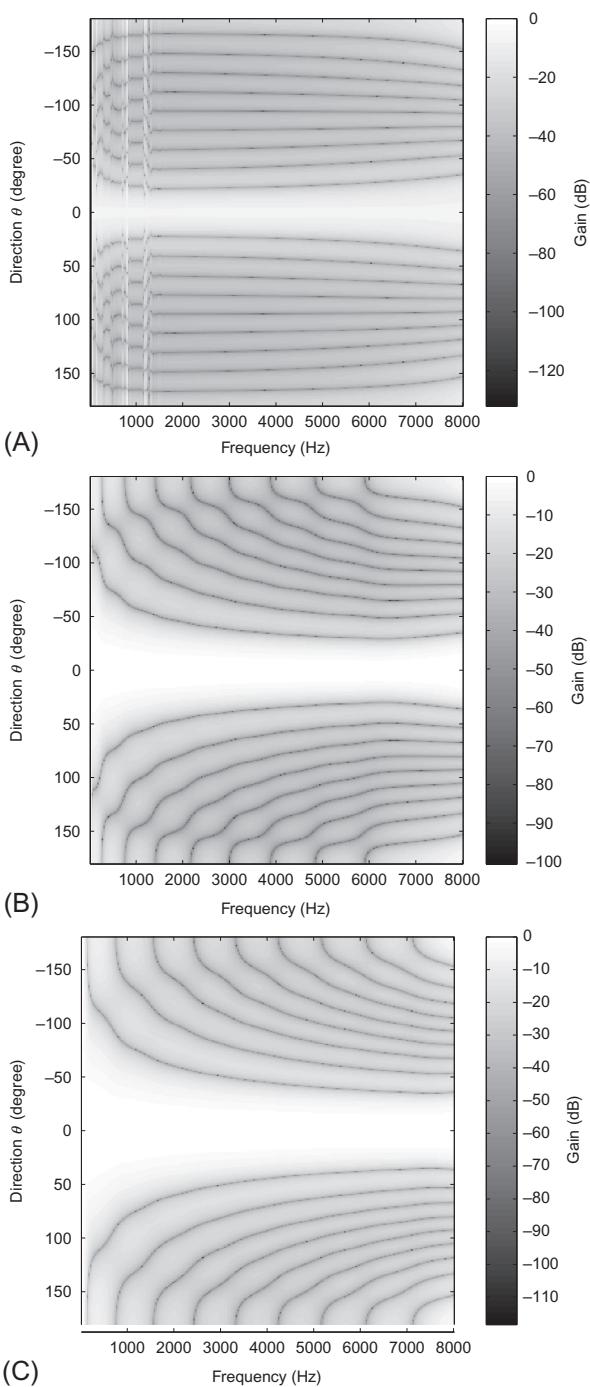
$$\mathbf{W}(k) = \frac{[\Gamma_V(k) + \epsilon \mathbf{I}_M]^{-1}\mathbf{A}_0(k)}{\mathbf{A}_0^H(k)[\Gamma_V(k) + \epsilon \mathbf{I}_M]^{-1}\mathbf{A}_0(k)}. \quad (12.54)$$

Comparing Eq. (12.53) with Eq. (12.54), it is clear that Eq. (12.54) is a regularized version of Eq. (12.53), where  $\epsilon$  can be seen as the regularization parameter. This parameter offers tradeoff between directivity and white noise amplification. A small  $\epsilon$  leads to a large DF and a low WNG, while a large  $\epsilon$  gives rise to a low DF and a large WNG. For comparison, the beampatterns of superdirective beamformers using a linear array of  $M = 10$ ,  $d = 2$  cm and  $\theta_0 = 0^\circ$  for  $\epsilon = 0$ ,  $\epsilon = 0.001$ , and  $\epsilon = 0.1$  are shown in Fig. 12.8. It may be noted that when  $\epsilon = 0$  the beamformer is the unregularized superdirective beamformer.

The weights of the fixed-beamformers discussed above are estimated based on a given source direction and geometry of the microphone array. The limitation of such beamformers is that the performance is limited by the number of microphones and the array geometry. However, it has been shown that by incorporating the signal statistics into the design of the beamformer it is possible to achieve better performance [6]. In the following sections we briefly describe one of the well-known data-dependent beamforming algorithms namely the linearly constrained minimum variance (LCMV) beamformer.

### 12.3.2 LINEARLY CONSTRAINED MINIMUM VARIANCE (LCMV)-BASED ADAPTIVE BEAMFORMING TECHNIQUES

The LCMV beamformer is considered as a general form of the adaptive beamformer. For ease of explanation, as in [11], we assume that the desired source signal impinges on the array from the broadside. With reference to Fig. 12.5, the LCMV algorithm

**FIG. 12.8**

The beampattern of superdirective beamformer using a linear array of  $M = 10$  microphones with  $d = 2$  cm and  $\theta_0 = 0^\circ$  with (A)  $\epsilon = 0$ , (B)  $\epsilon = 0.001$ , and (C)  $\epsilon = 0.1$ .

imposes constraints on the output of the beamformer in such a way that the signal impinging on the broadside of the array is left either undistorted or with a predetermined response while the power of the signals from all the other directions are minimized. Since the desired signal impinges on the broadside of the array, it propagates through the filters in parallel without any difference in delay. Hence, for a signal emanating from the broadside of the array, the beamformer can be replaced by a single microphone with a  $J$ -tap FIR filter with weights being the sum of the weights across each column of the beamformer filters. In terms of beamformer design, this constraint can be expressed mathematically as

$$\mathbf{c}_j^T \mathbf{w} = \tilde{w}_j, \quad j = 0, 1, \dots, J-1, \quad (12.55)$$

where  $\mathbf{c}_j = [\underbrace{0, \dots, 0}_M, \underbrace{0, \dots, 0}_M, \underbrace{1, \dots, 1}_{\text{jth group of } M}, \underbrace{0, \dots, 0}_M, \underbrace{0, \dots, 0}_M]^T$  is a column vector with elements being either 0 or 1. When this column vector is transposed and multiplied with  $\mathbf{w}$ , the  $j$ th coefficients of the beamformer filters will be summed to obtain  $\tilde{w}_j$ , which corresponds to the  $j$ th coefficient of the equivalent FIR filter. Extending the same for all the  $J$ -filter taps we have

$$\mathbf{C}^T \mathbf{w} = \tilde{\mathbf{w}}, \quad \mathbf{C} = [\mathbf{c}_0, \mathbf{c}_1, \dots, \mathbf{c}_{J-1}], \quad \tilde{\mathbf{w}} = [\tilde{w}_0, \tilde{w}_1, \dots, \tilde{w}_{J-1}]. \quad (12.56)$$

This will ensure that the signal from the array broadside will be constrained to a response determined by  $\tilde{\mathbf{w}}$ . Defining the autocorrelation matrix of  $\mathbf{x}(t)$  as

$$\mathbf{R}_{xx} = E[\mathbf{x}(t)\mathbf{x}^T(t)] \quad (12.57)$$

and if one were to minimize the variance of the beamformer output

$$\begin{aligned} E[y^2(t)] &= E[\mathbf{w}^T \mathbf{x}(t) \mathbf{x}^T(t) \mathbf{w}] \\ &= \mathbf{w}^T \mathbf{R}_{xx} \mathbf{w}, \end{aligned} \quad (12.58)$$

under the above constraint, suppression of signals from all the directions except from the broadside will be achieved. With the above, the LCMV problem can then be mathematically expressed as

$$\begin{array}{ll} \underset{\mathbf{w}}{\text{minimize}} & \mathbf{w}^T \mathbf{R}_{xx} \mathbf{w} \\ \text{subject to} & \mathbf{C}^T \mathbf{w} = \tilde{\mathbf{w}}, \end{array} \quad (12.59)$$

such that when solved using the Lagrange multiplier method, we obtain the LCMV solution as [11]

$$\mathbf{w} = \mathbf{R}_{xx}^{-1} \mathbf{C} [\mathbf{C}^T \mathbf{R}_{xx}^{-1} \mathbf{C}]^{-1} \tilde{\mathbf{w}}. \quad (12.60)$$

It may be noted that the optimum solution of  $\mathbf{w}$  requires prior knowledge of  $\mathbf{R}_{xx}$ . Since  $\mathbf{R}_{xx}$  varies with time, it requires frequent updating under practical scenarios.

In addition, we note that the inversion of  $\mathbf{R}_{xx}$  is required and to address this problem, Frost proposed an iterative algorithm for the estimation of  $\mathbf{w}$  as [11]

$$\mathbf{w}(t+1) = \left[ \mathbf{I} - \mathbf{C}(\mathbf{C}^T \mathbf{C})^{-1} \mathbf{C}^T \right] [\mathbf{w}(t) - \mu y(t) \mathbf{x}(t)] + \mathbf{C}(\mathbf{C}^T \mathbf{C})^{-1} \tilde{\mathbf{w}}, \quad (12.61)$$

where  $\mu$  is the adaptation step size.

In the above discussion, it was assumed that the desired signal impinges on the broadside of the array or is presteered. Presteering can be done either by mechanically rotating the array or by providing appropriate delays at the microphone outputs. However, the above formulation can be extended for signals from an arbitrary direction  $\theta$  and angular frequency  $\omega$  using the beamformer response in Eq. (12.26). This results in the following minimization problem

$$\begin{aligned} & \underset{\mathbf{w}}{\text{minimize}} \quad \mathbf{w}^T \mathbf{R}_{xx} \mathbf{w} \\ & \text{subject to } \mathbf{a}^H(\omega, \theta) \mathbf{w} = P^*, \end{aligned} \quad (12.62)$$

where  $P^*$  is a complex constant. Solving Eq. (12.62) using the Lagrange multiplier method, we have

$$\mathbf{w} = \frac{\mathbf{R}_{xx}^{-1} \mathbf{a}(\omega, \theta)}{\mathbf{a}^H(\omega, \theta) \mathbf{R}_{xx}^{-1} \mathbf{a}(\omega, \theta)} P^*. \quad (12.63)$$

When  $P^* = 1$ , we note that the beamformer will pass the desired signal without any distortion to the output and the beamformer is known as minimum variance distortionless response (MVDR) beamformer [6].

To reduce the computational complexity, the above formulation of LCMV beamformer can easily be implemented in the frequency domain as

$$\begin{aligned} & \underset{\mathbf{W}(k, l)}{\text{minimize}} \quad \mathbf{W}^H(k, l) \mathbf{R}_{XX}(k, l) \mathbf{W}(k, l) \\ & \text{subject to } \mathcal{A}^H(k) \mathbf{W}(k, l) = \mathbf{P}(k), \end{aligned} \quad (12.64)$$

where  $\mathbf{R}_{XX}(k, l) = E_l[\mathbf{X}^H(k, l) \mathbf{X}(k, l)]$ ,  $\mathcal{A}(k) = [\mathbf{A}_0(k), \mathbf{A}_1(k), \dots, \mathbf{A}_{I_d-1}(k)]$  for  $I_d$  constraints and  $\mathbf{P}(k) = [P^*_0(k), P^*_1(k), \dots, P^*_{I_d-1}(k)]$ . The optimum solution for the LCMV problem in Eq. (12.64) can be obtained using the Lagrange multiplier method to yield

$$\mathbf{W}(k, l) = \frac{\mathbf{R}_{XX}^{-1}(k, l) \mathcal{A}(k) \mathbf{P}(k)}{\mathcal{A}^H(k) \mathbf{R}_{XX}^{-1}(k, l) \mathcal{A}(k)}. \quad (12.65)$$

### 12.3.3 PRACTICAL CONSIDERATIONS IN COVARIANCE MATRIX ESTIMATION IN LCMV-BASED BEAMFORMERS

From the discussion in the previous sections we note that one of the important parameters required for the estimation of the beamformer weights is the covariance matrix  $\mathbf{R}_{XX}$ . If the environment is anechoic and the sources are in the far-field,  $\mathbf{R}_{XX}$  can be estimated directly from the microphone outputs. However, if the environment is

reverberant and the source-sensor separation is small, the response from source to the sensors cannot be modeled using the steering vectors alone. Under such scenarios,  $\mathbf{R}_{XX}$  must be estimated only from the noise and the interference signals components. This is particularly challenging under low SNR conditions. Much research has been done to address this problem [27, 28].

Recently, estimation of interference-plus-noise covariance matrix was achieved by exploiting the property of Hermitian angle in the TF domain [29]. In the STFT domain, in the absence of interferences and reflections, the microphone output can be expressed as

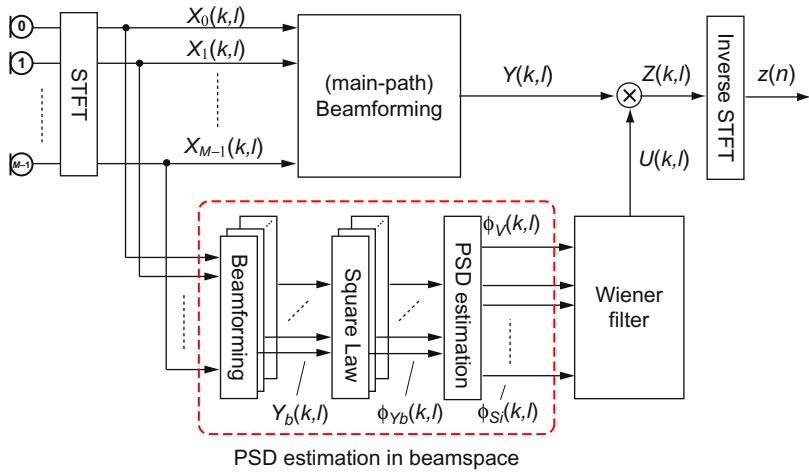
$$\mathbf{X}(k,l) = \mathbf{A}_0(k)S_0(k,l), \quad (12.66)$$

where  $S_0(k,l)$  is the desired source component in the STFT domain and  $\mathbf{A}_0(k)$  is the steering vector for the  $k$ th frequency bin. Eq. (12.66) shows that when only the desired signal component is present, i.e., in the absence of interferences and noise components,  $\mathbf{X}(k,l)$  is a scaled version of  $\mathbf{A}_0(k)$ . It can be shown that, using the property of Hermitian angle in complex vector space [30], the Hermitian angle between  $\mathbf{X}(k,l)$  and  $\mathbf{A}_0(k)$  will be zero. However, in the presence of other signals the Hermitian angle will be nonzero. It may be noted that in a practical scenarios it is difficult to find points with Hermitian angle being exactly zero due to reverberation and sensor noise. However, the technique can be used to find the probability of presence or absence of interference and noise components in each time-frequency point  $\mathbf{X}(k,l)$ . This can then be used for interference-plus-noise covariance matrix estimation.

## 12.4 POSTFILTER BY PSD ESTIMATION IN BEAMSPACE

Although beamforming can effectively separate a target sound source, its performance is often reduced in practical environment due to various causes of errors. These errors are often caused by modeling errors which include variation of microphones' sensitivities. Thus, a further measure to reduce the residual interfering signal is often taken. Of those applying a postfilter to the output of beamformer [31] is known to be effective to boost the performance of sound source separation. However, the power spectral densities (PSD) of both the target and interfering sound sources need to be accurately estimated to design an effective postfilter. In this section a method for estimating the PSD of sound sources using the combinations of directivity gains of multiple beamformers, also known as *PSD estimation in beamspace* [32], is introduced.

The signal flow of the beamformer with postfiltering, in which the PSD estimation in beamspace is embedded, is summarized in Fig. 12.9. The beamforming with postfiltering consists of a (main-path) beamformer, the Wiener postfilter, and a PSD estimation process. The PSD estimation in beamspace algorithm estimates PSDs for the Wiener filter calculation, which requires  $B$  beamformers besides the main-path beamformer. The details of the proposed PSD estimation algorithm are presented in the rest of this section.

**FIG. 12.9**

Source separation by beamforming and postfiltering.

### 12.4.1 PROBLEM SETUP

An  $M$ -sensor microphone array observes signals arriving from the  $N$  sound sources located at different angles  $\theta_i$  in a noisy environment as depicted in Fig. 12.10. Assume that the aperture of the microphone array is small enough compared to the distance to the sound sources so that the plane wave model [33] can be assumed for the observation. Let  $X_m(k, l)$  be the short-time Fourier transform of the signal observed by the  $m$ th microphone of the array where  $k$  and  $l$  denote the frequency and temporal frame, respectively.  $X_m(k, l)$  is represented by

$$X_m(k, l) = \sum_{i=1}^N H_{m, \theta_i}(k) S_i(k, l) + V_m(k, l), \quad (12.67)$$

where  $H_{m, \theta_i}(k)$  denotes the transfer function of the signal propagation path from the angle  $\theta$  to the microphone  $m$ , and  $S_i(k, l)$  and  $V_m(k, l)$  are the spectra of the sound source  $i$  and the incoherent noise component at the microphone  $m$ , respectively.

Each of the signals included in Eq. (12.67) is assumed to be uncorrelated with each other, i.e., the following relationships hold:

$$E[S_i(k, l) S_{i'}(k, l)] = 0 \quad i \neq i', \quad (12.68)$$

$$E[S_i(k, l) V_m(k, l)] = 0, \quad \forall i, \forall m. \quad (12.69)$$

Various models may be applied to the incoherent noise component  $V_m(k, l)$ . One of the models commonly used in previous studies [34] describes the incoherent noise component using its spatial propagation properties as

$$V_m(k, l) = \int_{\theta} H_{m, \theta}(k) V_{\theta}(k, l) d\theta, \quad (12.70)$$

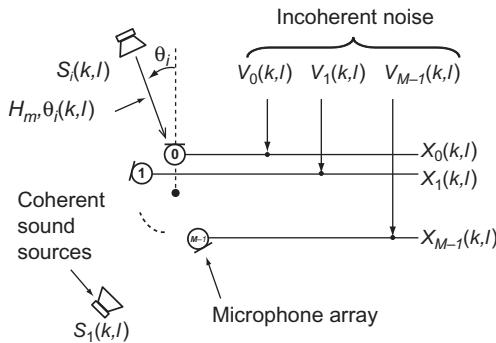


FIG. 12.10

Sound observation using microphone array in noisy environment.

where  $V_\theta(k, l)$  is the noise component arriving from the angle  $\theta$ . The model then applies the diffuse sound field model [35] to the incoherent noise where an isotropic distribution of the sound energy can be assumed. Introduction of this model replaces  $V_m(k, l)$  by

$$V_m(k, l) = V_{\bar{\theta}}(k, l) \int_{\theta} H_{m,\theta}(k) d\theta, \quad (12.71)$$

where  $V_{\bar{\theta}}(k, l)$  is the noise component averaged over the angle  $\theta$ .

Another model for the incoherent noise is the temporary stationary noise model where the power spectral density of the incoherent noise is assumed to be time invariant, which provides

$$V_m(k, l) = V_m(k). \quad (12.72)$$

Naturally these two models may hold simultaneously which makes the incoherent noise to be

$$V_m(k, l) = V_{\bar{\theta}}(k) \int_{\theta} H_{m,\theta}(k) d\theta. \quad (12.73)$$

The problem considered here aims at estimating the PSD of each coherent sound source  $S_i(k, l)$  and that of the incoherent noise  $V_m(k, l)$ .

### 12.4.2 BEAMFORMING AND ITS OUTPUT PSD

Let  $B(> N)$  different beamformers be applied to the microphone array observation. Given that  $W_{b,m}(k)$  is the frequency dependent weight of an arbitrary beamformer  $b$  for the microphone  $m$ , the output of the beamformer  $b$  is

$$Y_b(k, l) = \sum_{m=0}^{M-1} W_{b,m}(k) X_m(k, l) \quad (12.74)$$

$$\begin{aligned}
&= \sum_{i=1}^N \sum_{m=0}^{M-1} W_{b,m}(k) H_{m,\theta_i}(k) S_i(k, l) \\
&\quad + \sum_{m=0}^{M-1} W_{b,m}(k) \int_{\theta} H_{m,\theta}(k) V_{\theta}(k, l) d\theta
\end{aligned} \tag{12.75}$$

$$= \sum_{i=1}^N D_{b,\theta_i}(k) S_i(k, l) + \int_{\theta} D_{b,\theta}(k) V_{\theta}(k, l) d\theta, \tag{12.76}$$

where  $D_{b,\theta}(k) := \sum_{m=0}^{M-1} W_{b,m}(k) H_{m,\theta}(k)$  is the directivity function of the beamformer  $b$  to the angle  $\theta$ .

The PSD of the beamformer output can be easily calculated using Eq. (12.77), which can be approximated by Eq. (12.78).

$$\phi_{Y_b}(k, l) = E[|Y_b(k, l)|^2] \tag{12.77}$$

$$\approx \sum_{i=1}^N |D_{b,\theta_i}(k)|^2 \phi_{S_i}(k, l) + \int_{\theta} |D_{b,\theta}(k)|^2 \phi_{V_{\theta}}(k, l) d\theta \tag{12.78}$$

Here  $\phi_{S_i}(k, l)$  and  $\phi_{V_{\theta}}(k, l)$  are the PSD of the coherent signals and noise component arriving from angle  $\theta$ , respectively. Note the assumptions in Eqs. (12.68), (12.69) are used for the approximation from Eqs. (12.77), (12.78). In Eq. (12.77) if the signals are nonstationary, the PSD  $\phi_{Y_b}(k, l)$  can be recursively estimated as

$$\phi_{Y_b}(k, l) = \alpha \phi_{Y_b}(k, l-1) + (1-\alpha) |Y_b(k, l)|^2, \tag{12.79}$$

where  $\alpha$  is the forgetting factor set to a value between 0 and 1. With the diffused sound field assumption [35], the PSD of the incoherent noise can be represented by a constant value

$$\phi_{V_{\theta}}(k, l) = \phi_{V_{\bar{\theta}}}(k, l) = \text{const. } \forall \theta, \tag{12.80}$$

which provides a further approximation

$$\phi_{Y_b}(k, l) \approx \sum_{i=1}^N |D_{b,\theta_i}(k)|^2 \phi_{S_i}(k, l) + \phi_{V_{\bar{\theta}}}(k, l) \int_{\theta} |D_{b,\theta}(k)|^2 d\theta. \tag{12.81}$$

Eq. (12.81) explicitly expresses that the PSD of the beamformers' output can be approximated by the summation of the source PSDs multiplied by the directivity gain of the beamformers.

### 12.4.3 PSD ESTIMATION IN BEAMSPACE

The output PSDs of the  $B$  different beamformers give the simultaneous equations

$$\underbrace{\begin{bmatrix} \phi_{Y_1} \\ \vdots \\ \phi_{Y_B} \end{bmatrix}}_{\Phi_Y(k, l)} = \underbrace{\begin{bmatrix} |D_{1,\theta_1}|^2 & \dots & |D_{1,\theta_N}|^2 & \int_{\theta} |D_{1,\theta}|^2 d\theta \\ \vdots & \ddots & \vdots & \vdots \\ |D_{B,\theta_1}|^2 & \dots & |D_{B,\theta_N}|^2 & \int_{\theta} |D_{B,\theta}|^2 d\theta \end{bmatrix}}_{\mathbf{D}(k)} \underbrace{\begin{bmatrix} \phi_{S_1} \\ \vdots \\ \phi_{S_N} \\ \phi_{V_{\bar{\theta}}} \end{bmatrix}}_{\Phi_{S+V}(k, l)}. \tag{12.82}$$

Note that  $k$  and  $l$  are omitted in Eq. (12.82) for the sake of brevity. The PSD of the coherent signals, i.e.,  $\phi_{S_i}(k, l)$  and that of the diffused incoherent noise, i.e.,  $\phi_{V_{\bar{\theta}}}(k, l)$  are estimated by solving the simultaneous equation using the least squares method as

$$\hat{\Phi}_{S+V}(k, l) = \begin{cases} \mathbf{D}^{-1}(k)\Phi_Y(k, l) & (B=N+1) \\ \mathbf{D}^+(k)\Phi_Y(k, l) & (B>N+1) \end{cases} \quad (12.83)$$

where  $^+$  and  $\hat{\cdot}$  represent the Moore-Penrose pseudo inverse and an estimated value, respectively.

When the assumption of diffuse incoherent noise is not made, the substitution in Eq. (12.80) cannot be applied so that the PSD of the incoherent noise has to be estimated separately since it can no longer be estimated by PSD estimation in beamspace. In such case the PSD of the incoherent noise should be estimated using other properties such as temporal stationarity and be subtracted from the estimated PSD given by applying the PSD estimation in beamspace to the rearranged simultaneous equations [36]

$$\underbrace{\begin{bmatrix} \phi_{Y_1} \\ \vdots \\ \phi_{Y_B} \end{bmatrix}}_{\Phi_Y(k, l)} = \underbrace{\begin{bmatrix} |D_{1,\theta_1}|^2 & \dots & |D_{1,\theta_N}|^2 \\ \vdots & \ddots & \vdots \\ |D_{B,\theta_1}|^2 & \dots & |D_{B,\theta_N}|^2 \end{bmatrix}}_{\mathbf{D}'(k)} \underbrace{\begin{bmatrix} \phi_{S_1} \\ \vdots \\ \phi_{S_N} \end{bmatrix}}_{\Phi_S(k, l)} \quad (12.84)$$

and its solution is given by

$$\hat{\Phi}_S(k, l) = \begin{cases} \mathbf{D}'^{-1}(k)\Phi_Y(k, l) & (B=N) \\ \mathbf{D}'^+(k)\Phi_Y(k, l) & (B>N) \end{cases} \quad (12.85)$$

The PSD of incoherent noise component is estimated by taking the minimum value of the estimated PSD during a time interval

$$\hat{\phi}_{V_i}(k, l) = \min \{\hat{\phi}_{S_i}(k, l)\} \quad (12.86)$$

and is subtracted from the estimated source PSD  $\hat{\phi}_{S_i}(k, l)$  to separate the PSD of the sound sources and that of the incoherent noise.

#### 12.4.4 POSTFILTERING FOR SOURCE SEPARATION

The postfiltering is realized by applying the Wiener filter calculated using the estimated PSDs to the output signal of the main-path beamforming  $Y(k, l)$

$$Z(k, l) = U(k, l)Y(k, l), \quad (12.87)$$

where the Wiener filter for separating the  $n$ th source is derived by

$$U(k, l) = \frac{\hat{\phi}_{S_i}(k, l)}{\sum_{i=1}^N \hat{\phi}_{S_i}(k, l) + \hat{\phi}_V(k, l)}. \quad (12.88)$$

For the PSD of incoherent noise in Eq. (12.88), either  $\hat{\phi}_{V_{\overline{\theta}}}(k, l)$  or  $\hat{\phi}_{V_i}(k, l)$  can be used depending on the model assumed for the incoherent noise. Finally the output signal of the source separation is produced by applying inverse short time Fourier transform to  $Z(k, l)$ .

---

## 12.5 CONCLUSIONS

Beamforming is a spatial filtering technique to enhance signals from a desired direction relative to a microphone array and suppress noise and interferences from other directions. In this chapter, we introduced the basic concepts in beamforming and described the fundamental approaches in wideband beamformer design for speech signals. Specifically, we described the basic data-dependent and data-independent beamformers. We also discussed some of the practical design considerations such as noise-plus-interference covariance matrix estimation in data-dependent beamforming. Generally, beamformers suffer from performance degradation due to modeling errors which include variation of microphones' sensitivities. Therefore as a further measure, postfilters are often employed. In this chapter, we have introduced a postfiltering technique based on PSD estimation in beamspace to address this problem.

---

## REFERENCES

- [1] H.L. Van Trees, Optimum Array Processing, Part IV of Detection, Estimation, and Modulation Theory, John Wiley & Sons, New York, 2004.
- [2] J. Li, P. Stoica, Robust Adaptive Beamforming, vol. 88, John Wiley & Sons, New York, 2005.
- [3] W. Liu, S. Weiss, Wideband Beamforming: Concepts and Techniques, 17, John Wiley & Sons, New York, 2010.
- [4] M. Brandstein, D. Ward, Microphone Arrays: Signal Processing Techniques and Applications, Springer Science & Business Media, Berlin, Heidelberg, 2001.
- [5] J.P. Dmochowski, J. Benesty, S. Affes, A generalized steered response power method for computationally viable source localization, *IEEE Trans. Audio Speech Lang. Process.* 15 (8) (2007) 2510–2526.
- [6] B.D. Van Veen, K.M. Buckley, Beamforming: a versatile approach to spatial filtering, *IEEE ASSP Mag.* 5 (2) (1988) 4–24.
- [7] J. Dmochowski, J. Benesty, S. Affes, Linearly constrained minimum variance source localization and spectral estimation, *IEEE Trans. Audio Speech Lang. Process.* 16 (8) (2008) 1490–1502.
- [8] S. Doclo, M. Moonen, Superdirective beamforming robust against microphone mismatch, *IEEE Trans. Audio Speech Lang. Process.* 15 (2) (2007) 617–631.
- [9] H. Cox, R.M. Zeskind, T. Kooij, Practical supergain, *IEEE Trans. Acoust. Speech Signal Process.* 34 (3) (1986) 393–398.
- [10] M. Crocco, A. Trucco, Design of robust superdirective arrays with a tunable tradeoff between directivity and frequency-invariance, *IEEE Trans. Signal Process.* 59 (5) (2011) 2169–2181.

- [11] O.L. Frost, An algorithm for linearly constrained adaptive array processing, *Proc. IEEE* 60 (8) (1972) 926–935.
- [12] K.M. Ahmed, R.J. Evans, An adaptive array processor with robustness and broad-band capabilities, *IEEE Trans. Antennas Propag.* 32 (9) (1984) 944–950.
- [13] J. Capon, High-resolution frequency-wavenumber spectrum analysis, *Proc. IEEE* 57 (8) (1969) 1408–1418.
- [14] E.A.P. Habets, J. Benesty, I. Cohen, S. Gannot, J. Dmochowski, New insights into the MVDR beamformer in room acoustics, *IEEE Trans. Audio Speech Lang. Process.* 18 (1) (2010) 158.
- [15] Y. Gu, A. Leshem, Robust adaptive beamforming based on interference covariance matrix reconstruction and steering vector estimation, *IEEE Trans. Signal Process.* 60 (7) (2012) 3881–3885.
- [16] John E. Piper, Beamforming Narrowband and Broadband Signals, Sonar Systems, Prof. Nikolai Kolev (Ed.), InTech, 2011, <https://doi.org/10.5772/18445>. Available from: <https://www.intechopen.com/books/sonar-systems/beamforming-narrowband-and-broadband-signals>.
- [17] H. Nyquist, Certain topics in telegraph transmission theory, *Proc. IEEE* 90 (2) (2002) 280–305.
- [18] A.V. Oppenheim, R.W. Schafer, J.R. Buck, et al., *Discrete-Time Signal Processing*, vol. 2, Prentice Hall, Englewood Cliffs, NJ, 1989.
- [19] Y. Zhao, W. Liu, R.J. Langley, Efficient design of frequency invariant beamformers with sensor delay-lines, in: *Proc. IEEE Sensor Array and Multichannel Signal Processing Workshop*, IEEE, New York, 2008, pp. 335–339.
- [20] Y. Zhao, W. Liu, R. Langley, A least squares approach to the design of frequency invariant beamformers, in: *European Signal Processing Conference*, IEEE, New York, 2009, pp. 844–848.
- [21] R. Mars, V. Reju, A.W. Khong, A frequency-invariant fixed beamformer for speech enhancement, in: *Asia-Pacific Signal and Information Processing Association, 2014 Annual Summit and Conference (APSIPA)*, IEEE, New York, 2014, pp. 1–6.
- [22] Y. Zhao, W. Liu, R. Langley, Application of the least squares approach to fixed beamformer design with frequency-invariant constraints, *IET Signal Process.* 5 (3) (2011) 281–291.
- [23] S. Doclo, M. Moonen, Design of broadband beamformers robust against gain and phase errors in the microphone array characteristics, *IEEE Trans. Signal Process.* 51 (10) (2003) 2511–2526.
- [24] H. Duan, B.P. Ng, C.M.S. See, J. Fang, Applications of the SRV constraint in broadband pattern synthesis, *Signal Process.* 88 (4) (2008) 1035–1045.
- [25] R.K. Cook, R. Waterhouse, R. Berendt, S. Edelman, M. Thompson Jr, Measurement of correlation coefficients in reverberant sound fields, *J. Acoust. Soc. Am.* 27 (6) (1955) 1072–1077.
- [26] H. Cox, R.M. Zeskind, M.M. Owen, Robust adaptive beamforming, *IEEE Trans. Acoust. Speech Signal Process.* 35 (10) (1987) 1365–1376.
- [27] A. Khabbazibasmenj, S.A. Vorobyov, A. Hassanien, Robust adaptive beamforming based on steering vector estimation with as little as possible prior information, *IEEE Trans. Signal Process.* 60 (6) (2012) 2974–2987.
- [28] R. Mallipeddi, J.P. Lie, S.G. Razul, P. Suganthan, C.M.S. See, Robust adaptive beamforming based on covariance matrix reconstruction for look direction mismatch, *Prog. Electromagn. Res. Lett.* 25 (2011) 37–46.

- [29] H. Shen, V.G. Reju, A.W. Khong, Speech enhancement via covariance estimation using Hermitian angle in adaptive beamforming, in: Proc. IEEE Int. Conf. Digital Signal Process, IEEE, New York, 2015, pp. 1196–1200.
- [30] V.G. Reju, S.N. Koh, I.Y. Soon, Underdetermined convolutive blind source separation via time-frequency masking, *IEEE Trans. Audio Speech Lang. Process.* 18 (1) (2010) 101–116.
- [31] R. Zelinski, A microphone array with adaptive post-filtering for noise reduction in reverberant rooms, in: Proc. Int. Conf. Acoust., Speech, Signal Process, vol. 5, 1988, pp. 2578–2581.
- [32] Y. Hioka, K. Furuya, K. Kobayashi, K. Niwa, Y. Haneda, Underdetermined sound source separation using power spectrum density estimated by combination of directivity gain, *IEEE Trans. Audio Speech Lang. Process.* 21 (6) (2013) 1240–1250.
- [33] D. Johnson, D. Dudgeon, *Array Signal Processing: Concepts and Techniques*, Prentice Hall, Englewood Cliffs, NJ, 1993.
- [34] I. McCowan, H. Bourlard, Microphone array post-filter based on noise field coherence, *IEEE Speech Audio Process.* 11 (6) (2003) 709–716.
- [35] H. Kuttruff, *Room Acoustics*, fifth ed., Applied Science Publishers LTD, London, 2009.
- [36] K. Niwa, Y. Hioka, K. Kobayashi, Post-filter design for speech enhancement in various noisy environments, in: Proc. Int. Workshop Acoust. Signal Enhancement (IWAENC), 2014, pp. 35–39.

# Index

Note: Page numbers followed by *f* indicate figures, and *t* indicate tables.

## A

- Adaptive beamshaping, 180–182
- Adaptive detection algorithms, 241–242
- Adaptive normalized matched filter (ANMF), 240–242
- Adaptive radar, 339–343, 340*f*
- Adaptive radar detection
  - CES distributions, 215–239
  - estimation theory under model misspecification
    - CMCRB, 207–210
    - MML estimator, 205–206
    - MSE, 206–207
    - MS-unbiased estimators and MCRB, 202–205
    - regular models, 201–202
  - examples, 210–214
  - hypothesis testing problem for target detection, 240–245
  - problem statement and motivations, 200
  - symbols and functions, 197
- Additive white Gaussian noise (AWGN), 166, 341–343
- Aggregate beampattern, 37–38, 38*f*
- Aliasing, spatial, 591
- Alternating direction method of multipliers (ADMM), 570–571
- Amplify-and-forward (AF) relaying scheme, 405
- Amplitude modulation (AM), 8, 20
- Angle-of-arrival (AoA), 180–182, 393
- ANM. *See* Atomic norm minimization (ANM)
- Antenna array, direction-of-arrival, 292–294*f*, 293
- Antenna switching (AS) receiver, 480
- Antenna systems, integrated radar, 318–319, 319–320*f*
- Arbitrary waveform generators (AWGs), 12–13, 29*f*
- Atomic norm method, DOA estimation
  - multiple snapshot case, 559–561
  - single snapshot case, 543–547
- Atomic norm minimization (ANM), 543–544, 548, 557, 561
  - EMaC and, 548–549
  - GLS and, 551–552, 566–568
  - SMV-based, 555–557
- Automotive consumer radars, 346–354

## B

- Automotive radar, 317, 339–340
    - stochastic geometry technique, 346–354, 347*f*
    - interference analysis, 349–350
    - interference statistics, 350
    - lattice model, 348–349
    - performance analysis and optimization, 350–354, 351–352*f*, 354*f*
    - Poisson point process model, 347–348, 348*f*
    - ranging success probability, 346, 350–351, 351*f*
    - spatial success probability, 352–353, 352*f*
    - trends and standardization efforts, 321–322
  - Average case analysis, 524
  - Average sample number (ASN), 169
  - AWGN. *See* Additive White Gaussian Noise (AWGN)
- Bandwidth, limited system, 277–278
  - Bangs formula, 246–247
  - Barker code, modulation, 302
  - Barker-sidelobe-reduction (BSLR) filter, 303–304
  - Basis pursuit (BP), 516
  - Basis pursuit denoising (BPDN) problem, 517–518, 524, 535
  - Beamforming, 403–404, 587
    - end-to-end channel modeling, 404–405
    - massive MIMO, 369*f*
    - microphone array
      - DSB (*see* Delay-and-sum beamforming (DSB))
        - filter-and-sum beamforming, 587, 594–596, 594*f*
        - narrowband, 588–593
        - PSD estimation, 605–610, 606*f*
        - wideband, 593–605
      - one-way network, 405–406
        - frequency-flat channels, 406–422, 407*f*, 417*f*
        - frequency-selective channels, 422–427
        - numerical examples, 467–468, 468*f*
      - receive and transmit, 403
      - relay weight, 405, 432, 452, 455–456, 465
      - SWIPT, 482–502
        - co-located IR and ER, 483*f*, 497–502
        - IRs and ERs separation, 483*f*, 484–491

- Beamforming (*Continued*)  
     MISO system, 482, 483f  
     secret information, 491–497, 496f  
     two-way network, 427  
         asynchronous networks, 440–464, 451f, 462f  
         geometry, 469f  
         numerical examples, 468–472, 473f  
         sum-rate curves *vs.* total transmit power, 471–472f  
         synchronous networks, 428–440  
         with two transceivers, 429f
- Beamspace, PSD estimation, 605–610
- Beat signal, 330–331
- Bisection method, 447–448
- Bit error rate (BER)  
     minimization approach, 412  
     *vs.* total available transmit power, 470, 470f
- Bivector, 52, 73  
     attributes, 73f  
     geometric view of, 52f
- Blind methods, massive MIMO, 394
- Blocking matrix, 147
- BPDN problem. *See* Basis pursuit denoising (BPDN) problem
- Broadcast channel, 481–482, 489, 498
- BSLR filter. *See* Barker-sidelobe-reduction (BSLR) filter
- C**
- Capon's beamformer, 509, 566, 568
- Central limit theorem (CLT), 422
- CEP. *See* Constant envelop precoding (CEP)
- Channel capacity, 328–329, 356–357f, 357–359
- Channel-hardening behavior, 381–382
- Channel impulse response (CIR), 325–326, 404–405  
     end-to-end channel model, 442  
     relay channel, 440–441  
     two-way relaying scheme, 443–445  
     zero/nonzero taps, 446–448, 454, 458–459
- Channel state information (CSI), 367  
     acquisition, 367–368, 384  
     downlink, 384  
     estimation, 384–386  
     partial, 421–422  
     perfect/imperfect, BS antennas with, 368  
     pilot contamination, 387–389  
     response *vs.* tap index  $n$ , 473f  
     robust designs against errors, 422  
     SNR maximization, 408–410  
     at transmitter, 501, 503–504  
     user interference, 420–422
- Chirp, 329–330  
     acceleration codes, 22  
     quadratic phase function, 4  
     up-chirp/down-chirp, 5
- Chirp-rate-change codes, 22
- CIR. *See* Channel impulse response (CIR)
- Clutter, 255, 327–328
- CMBR. *See* Cosmic microwave background radiation (CMBR)
- Code-to-waveform (C2W) implementation, 14–15
- Cognitive radar, 155–156  
     adaptive waveform strategy, 156–157  
     all-digital apertures, 158  
     back-end computing, 158  
     canonical examples, 173–192  
         adaptive beamshaping, 180–182  
         carryover and adaptation performance gains, 182–185  
         detection performance, 176–178  
         information gained, 178–179  
         parallel estimation, 185–191  
         target detection with known impulse response, 174–179  
         target detection with known nuisance parameter, 179–185  
     early research contributions, 156–157  
     experiments, 192  
     hardware and processing technologies, enabling, 157–158  
     signal processing foundations, 159–173  
         POMDP, 172–173  
         sequential hypothesis testing, 168–172  
         waveform design, 159–168
- Coherence-inhibiting technique, DOA estimation, 532
- Coherent processing interval (CPI), 18–19
- Compact expression, 247–249
- Complementary Metal Oxide Semiconductors (CMOS)  
     applications, 339–340  
     improved platform for, 345f  
     millimeter-wave radar, 327–328  
     phase noise, 328  
     technology limitations, 354–355
- Complex elliptically symmetric (CES) distribution family  
     compact expression, 247–249  
     scatter matrix in, 199, 215–239
- Complex matrix, 103–113
- Complex number, 87–92
- Complex vector, 53–54  
     *N*-dimensional complex vector as 2*N*-dimensional real vector, 92–94

- examples, 100–103  
 observed data basis, 98  
 projecting into subspace, 98–103  
 as spinor expansion, 94–98  
 two-dimensional, 56  
**Compressed sensing (CS)**, 278  
 azimuth-range cell estimation, 299  
 estimation, 300f  
 group sparsity approach, 280  
 sparse representation and  
     convex relaxation, 516–518  
      $l_1$  optimization, 518–519  
     MLE, 519–520  
     problem formulation, 515–516  
**Consistent estimator**, 211  
**Constant envelop precoding (CEP)**, 380–381, 381f, 395  
**Constant false alarm rate (CFAR) detection**  
     process, 307  
**Constrained MCRB (CMCRB)**, 207–210, 233–235  
**Constrained MML (CMMML) estimator**, 231–233  
**Consumer radars, automotive**, 346–354, 347f  
 interference analysis, 349–350  
 interference statistics, 350  
 lattice model, 348–349  
 performance analysis and optimization,  
     350–354, 351–352f, 354f  
 Poisson point process model, 347–348, 348f  
 ranging success probability, 346, 350–351, 351f  
 spatial success probability, 352–353, 352f  
**Continuous phase modulation framework**, 20–21  
**Continuous wave (CW) waveform**, 340–342  
**Contraction rule**, 68  
**Convex relaxation**  
     on-grid sparse methods, 524–527  
     sparse representation and CS, 516–518  
**Coplanar vectors**, 87f  
**Co-prime arrays**, 555–556  
**Co-prime sampling**  
     in range, 283  
     slow-time, 283–284  
     spectral domain, 289–291  
**Corner reflectors**, 271–272  
**Cosmic microwave background radiation (CMBR)**, 357–359  
**Covariance fitting methods**  
     multiple snapshot case, 553–557  
     single snapshot case, 550–551  
**Covariance matrix estimation**, 604–605  
**Cramér-Rao lower bound (CRB)**, 198  
**CS. *See* Compressed sensing (CS)**  
**CSI. *See* Channel state information (CSI)**  
**Cumulative distribution function (CDF)**, 346, 350
- D**
- Data model, DOA estimation, 511  
 gridless sparse methods, 538  
 off-grid sparse methods, 533–534, 536  
 on-grid sparse methods, 522–523  
**Delay-and-sum beamforming (DSB)**, 587, 588f, 599–600  
 beampattern, 600f  
 and FIR filter, 592–594  
 with microphones, 591f  
 operating principle, 589–590  
 signals of different frequencies, 590–591, 590f  
 wideband, 600f, 601  
**Delay-Doppler ambiguity function**, 7, 7f  
**Derivative phase shift keying (DPSK)**  
     binary codes implementation with, 18–19, 18f  
     implementation, 15f  
**Detection algorithms**, 257–258  
**Deterministic sparse methods**  
     atomic  $l_0$  norm, 558–559  
     atomic norm, 559–561  
     framework for, 542, 557–558  
     Hankel-based nuclear norm, 561–562  
     multiple snapshot case, 557–562  
**DFT. *See* Discrete Fourier transform (DFT)**  
**Differential SAR tomography**, 295  
**Dimensionality reduction technique**, 525–527, 569–570  
**Direct digital synthesizers**, 12–13  
**Direction-of-arrival (DOA) estimation**, 291–293, 509–511  
 antenna array, 292f  
 circular antenna array, 293, 294f  
 data model, 511  
 gridless sparse methods, 510, 537–538  
     ANM vs. GLS, 566–568  
     computational issues and solutions, 568–571  
     data model, 538  
     multiple snapshot case, 553–562  
     reweighted atomic norm minimization, 562–566  
     single snapshot case, 541–552  
 Vandermonde decomposition of Toeplitz covariance matrices, 537–541  
 linear antenna array, 293, 293f  
 off-grid sparse methods, 510  
     dynamic grid, 536–537  
     fixed grid, 533–536  
 on-grid sparse methods, 510, 522  
     coherence-inhibiting technique, 532  
     convex relaxation, 524–527  
     data model, 522–523  
     dimensionality reduction technique, 525–527

- Direction-of-arrival (DOA) estimation (*Continued*)
   
    GLS, 528–529
   
     $l_2, 0$  optimization, 523–524
   
     $l_2, q$  optimization, 527–528
   
    maximum likelihood estimation, 531–532
   
    selection, 532–533
   
    singular value decomposition, 525–527
   
    SPICE, 529–531
   
    parameter identifiability, 513–515
   
    problem, 198–199
   
    research challenges, 571–572
   
    role of array geometry, 511–513
   
    sparse representation
   
        and CS, 515–520
   
        link and gap, 521–522
   
    2D array, 293, 511–513, 512*f*
  
    uniform linear array, 512
- Directivity factor (DF), 601
- Discrete Fourier transform (DFT)
   
    multicarrier equalization, 443
   
    normalized, 342
   
    two-dimensional, 343
- DOA estimation. *See* Direction-of-arrival (DOA) estimation
- Doppler frequency, 12
   
    detection of significant, 341–342
   
    sparsity sampling in, 283–286, 286*f*
- Doppler resolution, 328–329
- Doppler tolerance, 7
- Dynamic grid, DOA estimation
   
    data model, 536
   
    EM algorithm, 536–537
   
    gradient descent method, 537
- E**
- EGOMP. *See* Extended Group Orthogonal Matching Pursuit (EGOMP)
- Eigenbivectors, 107
- EIRP. *See* Equivalent isotropic radiated power (EIRP)
- Electromagnetic (EM) field, 354–357
- Electromagnetic (EM) ray tracing, 324–325
- EM algorithm. *See* Expectation-maximization (EM) algorithm
- End-to-end channel model
   
    beamforming, 404–405
   
    two-way network beamforming, 442
- Energy beamforming, SWIPT, 482–502
- Energy efficiency (EE), 368, 410–411
- Energy harvesting, 479, 482, 484
- Energy spectral density (ESD), 165
- Energy spectral variance (ESV), 165
- Enhanced matrix completion (EMaC), 548–549
- Equivalent isotropic radiated power (EIRP), 322
- Estimation of parameters by rotational invariant techniques (ESPRIT), 509, 540–541
- Estimation theory
   
    CMCRB, 207–210
   
    MML estimator, 205–206
   
    MSE, 206–207
   
    MS-unbiased estimators and MCRB, 202–205
   
    regular models, 201–202
   
    Euler's formula, 53
   
    Expectation-maximization (EM) algorithm, 520, 532
   
    off-grid sparse methods, 536–537
- Extended Group Orthogonal Matching Pursuit (EGOMP), 309, 312*f*
- F**
- FEKO, 270–272, 273*f*
- Fiber-based Ethernet networks, 263
- Field programmable gate array (FPGA) technology, 256
- Filter-and-forward (FF) relaying scheme, 405, 427, 441
- FIR filter, 422–424
- frequency-selective transceiver-relay links, 466–467
- Filter-and-sum beamforming, 587, 594–596, 594*f*
- Finite impulse response (FIR) filter, 588, 592–594
- end-to-end channel modeling, 404–405
- FF protocol, 422–424, 466–467
- J*-tap, 601–604
- Fisher information matrix (FIM), 202
- Fixed beamformer, 144–145
- Fixed grid, 533–536
   
    data model, 533–534
   
     $l_1$  optimization, 534–535
   
    sparse Bayesian learning, 536
- Flicker noise, 344–346, 344*f*
- FMCW. *See* Frequency modulated continuous wave (FMCW)
- Focal underdetermined system solver (FOCUSS), 519, 527
- Folded clutter, 19
- Four-dimensional real vector, 55
- Fourier transform, 25
- Fractional bandwidth, 39
- Frequency-change (chirp rate), 22
- Frequency-division duplexing (FDD), 367–368
- Frequency-flat channels, 406
   
    multiuser networks, 416–417, 417*f*
  
    orthogonal user channels, 417–420

- robust designs against CSI errors, 422  
 user interference and partial CSI, 421–422  
 user interference and perfect CSI, 420–421  
 single-user networks, 406–407, 407f  
   partial CSI, 412–414  
   SNR-maximization with perfect CSI, 408–410  
   SNR-per-unit-power maximization, 410–411  
 Frequency-invariant beamformer (FIB), 596–597, 597f  
 Frequency modulated continuous wave (FMCW), 329–331, 331f  
 Frequency modulated (FM) waveforms, 4–9, 20  
 Frequency template error (FTE), 11
- G**
- Gaussian PSD, 10–11  
   in dB, 11f  
   oversampling, 25  
 Gaussian spectral mask, 22–23  
 Generalized least squares (GLS) method, 557  
   ANM vs., 551–552, 566–568  
   on-grid sparse methods, 528–529  
 Generalized likelihood ratio test (GLRT), 198–199  
 Generalized Lloyd algorithm, 412  
 Generalized sidelobe canceller (GSLC), 147  
 General likelihood ratio test (GLRT), 298  
 Geometric algebra  
   complex matrix, 103–113  
   complex number, 87–92  
   complex vector, 92–103  
   multivectors, 82–87  
   pseudoscalar, 81–82  
   in three dimensions, 77–81  
   in two dimensions, 74–77  
   vector multiplication, 57–74  
 Geometry, 51  
   formulating detectors, 124–127  
   of matrix inverse, 109–113  
   of multistatic radar system, 260f  
   notch filter, 131–136  
   of nulling directions, 128–148  
   of signal detection, 119–127  
 Gerchberg-Saxton (GS) algorithm, 24  
 GLS method. *See* Generalized least squares (GLS) method  
 GPS Disciplined Oscillators (GPSDOs), 262–263  
 Grade-lowering operation, 83  
 Grade-raising operation, 83  
 Gradient-based algorithm, 426, 436  
 Gradient descent method, 537  
 Gridless sparse methods, DOA estimation, 510, 537–538  
 ADMM, 570–571  
 ANM vs. GLS, 566–568  
 computational issues and solutions, 568–571  
 data model, 538  
 dimensionality reduction technique, 569–570  
 multiple snapshot case  
   covariance fitting methods, 553–557  
   deterministic methods, 557–562  
 reweighted atomic norm minimization, 562–566  
 single snapshot case, 541–542  
   ANM vs. GLS, 551–552  
   atomic  $l_0$  norm, 542  
   atomic norm, 543–547  
   covariance fitting method, 550–551  
   deterministic sparse methods, 542  
   Hankel-based nuclear norm, 547–548  
 Vandermonde decomposition of Toeplitz  
   covariance matrices, 537–541  
 Group sparsity, 300–311, 301f  
 group model, 302  
 MIMO radar network, 300, 305–306, 305–306f  
 SFN radar, 306–311, 307–308f  
   EGOMP method, 309, 312f  
   signal model, 308–309  
   verification, 309–311, 310–312f  
 SIMO radar network, 302–304, 303–305f
- H**
- Hamming correlation, 335  
 Hankel-based nuclear norm, 547–548, 561–562  
 Hermitian inner product, 113–119  
 High-power amplifier (HPA), 5–7, 12–13  
 Hilbert-Schmidt operator approximation  
   techniques, 328  
 Holistic higher-dimensional waveform diversity, 35–44  
 Holistic waveform, implementation and design, 19–35  
 Holistic wideband MIMO radar, 38–44  
 Hybrid FM, 8  
 Hyperbolic FM (HFM), 10  
 Hypothesis testing problem, for target detection, 240–245
- I**
- Imaginary space, 15  
 Infinite impulse response (IIR) filters, 426  
 Information theory, limitations, 355–359, 356–359f  
 Integrated radar  
   antenna systems, 318–319, 319–320f

- Integrated radar (*Continued*)  
 automotive, 321–322  
 interference challenges, 320–321  
 millimeter-wave, 327–328  
     automotive scenario, 324*f*  
     clutter, 327–328  
     impulse responses for time-varying scene, 327*f*  
     propagation properties, 322–323, 323*f*  
     radar equation, 323–324  
     ray tracing, 324–326, 326*f*  
 single chip RF system, 318, 318–319*f*  
 system design challenges, 317  
     trends and standardization efforts, 321–322  
 Integrated sidelobe level (ISL), 10, 10*f*  
 Interference cancellation method, 342  
 Interference-free (IF) system, 377–380  
 Intermediate frequency (IF) processing technique, 343–346  
 International Telecommunication Union (ITU), 322  
 Inter-symbol-interference (ISI), 423–424, 440–441, 465  
 Inverse discrete Fourier transform (IDFT), 331–332, 337  
 Inverse fast Fourier transform (IFFT) algorithm, 25, 290, 337  
 Inverse vector, 60  
 Invisible space, 15, 35
- J**  
 Joint prechannel and postchannel equalization, 461–464, 462*f*
- K**  
 Kalman tracking, 187  
 Klystrons, 5–7  
 Kronecker’s theorem, 549  
 Kullback-Leibler (KL) divergence, 198
- L**  
 Lagrange multiplier method, 412–413, 424–425  
     LCMV, 601–604  
     wideband beamforming, 595–597, 599–600  
 Large-scale antenna systems. *See* Massive multiinput multioutput (MIMO)  
 Lattice model, automotive consumer radars, 348–349  
 Least absolute shrinkage and selection operator (LASSO), 517–518, 524  
     square-root, 518, 531  
     SVD, 525–526
- Least squares (LS) method  
     frequency-invariant beamformer, 596–597  
     wideband beamforming, 595–596  
 Lévy distribution, 350  
 Light processing techniques, 339–343  
 Linear amplification using nonlinear components (LINC), 29, 29*f*  
 Linear FM (LFM), 4  
     additional benefits of, 7  
     conservation of ambiguity, 8  
     delay-Doppler ambiguity function, 7, 7*f*  
     matched filter responses for, 6*f*  
     power spectral densities of, 6*f*  
     radar, 329–331, 330*f*  
     time-frequency relationship for, 5*f*  
     waveform, transmitter distortion and, 13–14*f*  
 Linearly constraint minimum variance (LCMV) beamformer, 601–604  
     covariance matrix estimation, 604–605  
     Lagrange multiplier method, 601–604  
 Linear period modulation (LPM) waveforms, 10  
 Linear threshold detector (LTD), 241  
 Locally convergent iterative algorithm, 564–565  
 Low noise amplifiers (LNAs), 354–355
- M**  
 MABC. *See* Multiple access broadcast channel (MABC)  
 Majorization-minimization (MM) algorithm, 519–520, 532  
 Massive multiinput multioutput (MIMO), 35, 367–370  
     beamforming, 369*f*  
     channel estimation, 384–386  
      colocated, 15  
     energy consumption, 369  
     energy efficiency, 368  
     features, 368–370  
     imaginary space, 40  
     inherent complexity, 394–395  
     invisible space, 40  
     invisible space regions, 40, 40*f*  
     limiting cases, 372  
     millimeter wave band, 396  
     minimum mean-square-error detectors, 382, 385–386, 392–393  
     multiplexing gain, 368  
     MU-MIMO, 370, 373–376  
         downlink, 375–376  
         uplink, 374–375  
     peak-normalized near-omnidirectional optimization, 43*f*

- pilot contamination, 386–394, 390*f*  
 AoA-based methods, 393  
 blind methods, 394  
 mitigating effects, 392–394  
 precoding methods, 393  
 protocol-based methods, 393  
 scheduled UEs, 392*f*  
 point-to-point, 370–372, 371*f*  
 power control, 383–384  
 practical issue for, 15  
 precoding techniques, 376–381  
   AMP-based, 380  
   basic schemes, 377–380  
   CEP, 380–381, 381*f*, 395  
   linear and nonlinear, 376–377  
   maximum ratio transmission, 383–384  
   regularized ZF, 378  
   single-cell MU-MIMO, 379*f*  
   SNR/SINR expressions, 377*t*  
   zero forcing, 377–380  
 random matrix theory analysis, 372  
 research challenges, 394–396  
 robustness and reliability, 369  
 signal detection, 381–383  
 simple linear processing, 369  
 space-frequency response, 42*f*  
 spatial modulation, 394–395  
 spectral efficiency, 368  
 TDD system, 367–368, 375, 384–387, 385*f*  
 wideband radar, 38–44
- Matched filter (MF), 280, 303  
 group sparsity technique, 303  
 massive MIMO precoding techniques, 378–379  
 MU-MIMO, 375  
 nonparametric method, 300*f*  
 precoder, 375–376  
 pulse compression of chirped pulse, 282*f*  
 pulse-Doppler radar, 286*f*  
 SNR output, 328–329  
 sparse sampling, 282, 313  
 target positions determination, 306, 306*f*  
 temporal sparsity, 280–281
- Matrix inverse  
 geometry of, 109–113  
 observed data basis, 110
- Maximum likelihood estimation (MLE), 198, 509  
 advantage, 519–520  
 on-grid sparse methods, DOA estimation, 531–532  
 sparse representation, 519–520
- Max-min SNR fair design approach, 436–437,  
 439–440
- asynchronous networks, 443–450
- synchronous networks, 435–437
- MCRB. *See* Misspecified Cramér-Rao bound (MCRB)
- Mean-squared-error (MSE), 206–207  
 minimization problem, 414–416  
 total MSE minimization, 452–455
- MF. *See* Matched filter (MF)
- Micro-Doppler effect, multistatic radar systems, 255, 265–268, 267*f*
- Microphone array  
 beamforming  
   DSB (*see* Delay-and-sum beamforming (DSB))  
   filter-and-sum beamforming, 587, 594–596,  
     594*f*  
   narrowband, 588–593  
   PSD estimation, 605–610, 606*f*  
   wideband, 593–605
- M*-sensor, 606
- sound observation in noisy environment, 606–607, 607*f*
- Millimeter-wave radar, 327–328  
 automotive scenario, 324*f*  
 clutter, 327–328  
 impulse responses for time-varying scene, 327*f*  
 propagation properties, 322–323, 323*f*  
 radar equation, 323–324  
 ray tracing, 324–326, 326*f*
- Minimum mean-square-error (MMSE) detectors, 382, 385–386, 392–393
- Minimum shift keying (MSK), 8–9
- Minimum variance distortionless response (MVDR) beamformer, 604
- Mismatched maximum likelihood (MML) estimator, 198–199, 205–206  
 examples, 210–214
- Misspecified Cramér-Rao bound (MCRB), 198–199, 203  
 as bound on mean square error, 206–207  
 compact expression for, 247–249  
 constrained MCRB, 207–210  
 examples, 210–214  
 for intrinsic parameter vector, 208–210  
 MS-unbiased estimators and, 202–205  
 MS-unbiasedness and, 209–210  
 for scatter matrix estimation, 215–239
- Misspecified data model, 200
- MLE. *See* Maximum likelihood estimation (MLE)
- MM algorithm. *See* Majorization-minimization (MM) algorithm
- Model-in-the-Loop (MiLo) optimization, 32–33
- Monostatic radars, 262
- MS-unbiased estimators, and MCRB, 202–205
- MS-unbiasedness, and MCRB, 209–210

- Multiantenna SWIPT, 481–482  
 Multicarrier equalization, 443  
 Multiinput multioutput (MIMO), 277–278  
     group sparsity, 300, 305–306,  
     305–306f  
     massive (*see* Massive multiinput multioutput (MIMO))  
     radar, 253  
 Multiple access broadcast channel (MABC), 427,  
     429–430, 440  
 Multiple-input single-output (MISO)  
     broadcast channel, 482, 489, 498  
     SWIPT system, 482, 483f  
 Multiple measurement vectors (MMVs), 523  
 Multiple signal classification (MUSIC), 509,  
     556–557  
 Multiple snapshot case, DOA estimation  
     covariance fitting methods, 553–557  
     deterministic methods, 557–562  
 Multiple-time-around clutter, 19  
 Multiplexing gain, 368  
 Multipoint-to-multipoint SWIPT, 502–503  
 Multistatic radar systems, 262  
     characteristics, 253–255  
     corner reflectors FEKO simulation,  
         271–272  
     down-range profiles, intersection, 254f  
     enablers, 256  
     NetRAD, 263–268  
     NeXtRAD, 268–269  
     polarimetric radar, 269–271  
     signal processing in, 256–257  
     synchronization considerations  
         for, 262–263  
     target detection, 257–258  
     target localization, 259–262  
     target resolution, 258–259  
 Multiuser MIMO (MU-MIMO), 370, 373–376  
     downlink, 375–376  
     single-cell, 379f  
     uplink, 374–375  
 Multiuser networks  
     frequency-flat channels, 416–417, 417f  
     orthogonal user channels, 417–420  
     robust designs against CSI errors, 422  
     user interference and partial CSI, 421–422  
     user interference and perfect CSI, 420–421  
     frequency-selective channels, 427  
 Multivariate Gaussian PDF, 119–124  
 Multivectors, 67  
 MU-MIMO. *See* Multiuser MIMO (MU-MIMO)  
 Mutual coherence, 280, 516–517, 521  
 Mutual information (MI), 155, 178f, 183f,  
     371–372
- N**
- Narrowband beamforming, 588–593  
 NetRAD, 263–268  
 Network beamformer, 403. *See also* Beamforming  
 Networks with frequency-selective transceiver-  
     relay links, 464–467  
     filter-and-forward relaying, 466–467  
     OFDM-based channel equalization, 465–466  
 NeXtRAD, 268–269  
 Nonlinear FM (NLFM) forms  
     matched filter responses for, 6f  
     power spectral densities of, 6f  
 Nonlinear least squares (NLS) method, 509  
 Notch filter, designing, 131–136  
 Nuclear norm minimization (NNM), MUSIC,  
     556–557  
 Nulling directions  
     adaptive processor, 145f  
     frequency choosing that define constraint  
         subspace, 136–138  
     generalized sidelobe canceller, 138–148  
     geometry of, 128–148  
     linear processing to steer nulls, 128–130  
     notch filter, 131–136  
     unconstrained minimization problem, 146  
 Nyquist-Shannon criteria, 278, 280–281, 298
- O**
- OFDM. *See* Orthogonal frequency-division  
     multiplexing (OFDM)
- Off-grid sparse Bayesian inference (OGSBI), 536
- Off-grid sparse methods, DOA estimation, 510  
     dynamic grid  
         data model, 536  
         EM algorithm, 536–537  
         gradient descent method, 537  
     fixed grid, 533–536  
         data model, 533–534  
          $l_1$  optimization, 534–535  
         sparse Bayesian learning, 536
- One-way network beamforming, 405–406  
     frequency-flat channels, 406  
         multiuser networks, 416–417, 417f  
         single-user networks, 406–407, 407f  
     frequency-selective channels, 422–427  
         multiuser networks, 427  
         single-user networks, 423–426  
     numerical examples, 467–468, 468f
- On-grid sparse methods, DOA estimation, 510, 522  
     coherence-inhibiting technique, 532  
     convex relaxation, 524–527  
     data model, 522–523  
     dimensionality reduction technique, 525–527

GLS, 528–529  
 $l_2, 0$  optimization, 523–524  
 $l_2, q$  optimization, 527–528  
maximum likelihood estimation, 531–532  
multiple measurement vectors, 523  
selection, 532–533  
single measurement vectors, 523  
singular value decomposition, 525–527  
SPICE, 529–531

Optimal energy beamformer (OeBF), 487  
Oriented area, 52  
Oriented length, 52  
Oriented volume, 52  
Orthogonal frequency-division multiplexing (OFDM), 289, 302  
equalization schemes, 461  
two-way relay network, 444f  
Orthogonal projection, 86, 87f  
Orthonormal vectors, 103  
Overcoding, 23

## P

Parseval’s theorem, 446  
Partially observable Markov decision processes (POMDP), 172–173  
PCP method. *See* Pilot contamination precoding (PCP) method  
Peak sidelobe level (PSL), 9–10, 10f  
Peak sidelobe ratio, 9–10  
Peak-to-average power ratio, 9  
Percent power bandwidth, 40  
Phase code (PC), 8  
Phase noise, 328, 343  
Physical emission, 31  
Pilot contamination  
channel-state information, 387–389  
massive MIMO, 386–394, 390f  
AoA-based methods, 393  
blind methods, 394  
mitigating effects, 392–394  
precoding methods, 393  
protocol-based methods, 393  
scheduled UEs, 392f  
Pilot contamination precoding (PCP) method, 393  
Point-to-point MIMO, 370–372, 371f  
Poisson point process (PPP) model, 346–348, 348f  
Polarmetric radar, 269–271  
Polyphase-coded FM (PCFM), 19–23  
autocorrelation, 22–24f  
Power added efficiency (PAE), 31–32  
Power control, massive MIMO, 383–384  
Power spectral density (PSD)  
diffuse sound field model, 606–607

estimation in beampspace, 605–610  
Gaussian-shaped PSD, 10–11  
of LFM and NLFM waveforms, 5, 6f  
microphone array in noisy environment, 606–607, 607f  
output beamformer, 607–608  
of output signal contribution, 163–164  
postfiltering for source separation, 609–610  
symmetric Gaussian random process with, 160–161

Toeplitz matrix, 538–539, 541, 550  
Power spectral variance (PSV), 165  
Power-splitting (PS) receiver, SWIPT, 480, 480f  
PPP model. *See* Poisson point process (PPP) model  
Precision Time Protocol (PTP), 263  
Precoding techniques, massive MIMO, 376–381  
AMP-based, 380  
basic schemes, 377–380  
CEP, 380–381, 381f, 395  
linear and nonlinear, 376–377  
maximum ratio transmission, 383–384  
regularized ZF, 378  
single-cell MU-MIMO, 379f  
SNR/SINR expressions, 377t  
zero forcing, 377–380

Probability density function (PDF), 119–124, 168–169

Probability of success  
ranging, 346, 350–351, 351f  
spatial, 352–353, 352f

Protocol-based methods, massive MIMO, 393

Pseudo-random stepped frequency (PRSF)  
waveform, 333–336, 334f, 339f  
continuous wave and, 340–341  
signal processing, 336–339, 343

Pseudoscalar, 80–82

Pseudo-true parameter vector, 198

Pulse agility, 18–19

Pulse compression, 4–12

BSLR filter, 303  
chirped pulse, 282f  
matched filter, 282f  
sensing matrix depends on, 303–304

Pulse Doppler radar, 283, 285  
nonuniform sampling, 286f  
temporal sparsity, 280–281  
uniform sampling, 286f

Pulsed radars, 16–17, 16f

## Q

Quadratically constrained quadratic programs (QCQPs), 486–488  
Quasi maximum likelihood (QML) estimator, 198

**R**

Radar cross section (RCS), 255, 271  
 Radar emissions, transmitter-in-the-loop  
     optimization of, 32*f*  
 Radar equation, 323–324  
 Radar interference  
     analysis, 349–350  
     challenges, 320–321  
     statistics, 350  
 Radar modes, 15  
 Radar signal processing, applications  
     Hermitian inner product, 113–119  
     nulling directions, 128–148  
     signal detection, 119–127  
 Radar system, 153–156  
     adaptive, 339–343, 340*f*  
     conventional methods for designing software,  
         344*f*  
     integrated (*see* Integrated radar)  
 PRSF, 333–339, 334*f*, 339*f*  
 resolution, 328–329  
 sparsity reconstruction (*see* Sparse  
     reconstruction techniques)  
 stepped frequency, 331–332, 332*f*  
 waveform and signal processing, 328–346  
 Radar waveform, 3–9  
 Random matrix theory analysis, massive MIMO,  
     372  
 Random stepped frequency (RSF), 321, 332, 343,  
     345*f*  
 Range  
     correlation effect, 328  
     resolution, 328–329  
     sidelobe modulation effect, 18–19  
     sparsity sampling in, 283–286, 286*f*  
     straddling, 17  
 Range-ambiguous clutter, 19  
 Ranging success probability, 346, 350–351, 351*f*  
 Ray tracing, millimeter-wave radar, 324–326, 326*f*  
 Received signal structure, 12  
 Receive effects, 16–19  
 Reference matrix, 337–338  
 Regularized ZF (RZF) precoding techniques, 378  
 Restricted isometry constant (RIC), 517  
 Restricted isometry property (RIP), 278–279, 517,  
     521  
 Reweighted atomic norm minimization (RAM),  
     562–566  
 Robust range-Doppler estimation, 342–343  
 Root of the MSE (RMSE), 213  
 Rotors, 91, 92*f*  
 RSF. *See* Random stepped frequency (RSF)

**S**

Sample covariance matrix (SCM), 217  
 Sample mean estimator, 210  
 SBL. *See* Sparse Bayesian learning (SBL)  
 Scalloping, 17  
 Scatter matrix estimation  
     complex Normal model for data, 222–224  
     generalized Gaussian, 224–230  
     MCRB for, 215–239  
     misspecified estimation, 216–230  
     misspecified joint estimation, 230–239  
     performance analysis, 235–239  
 Secondary surveillance radar (SSR), 254–255  
 Secrecy beamforming design, SWIPT, 491  
     numerical results, 495–497, 495*f*  
     optimal beamforming solution, 493–494  
     problem formulation, 493  
     system model, 491–493  
 Self-interference, 427, 429–430, 461  
 Semidefinite relaxation (SDR), 427, 487–488  
 Sequential hypothesis testing, 168–172  
 Sequential probability ratio test (SPRT),  
     156–157  
 Short-term frequency, 262  
 Short-time Fourier transform (STFT), 593,  
     597–598, 605  
 Sidelobe canceller, 138–148  
 Signal detection  
     formulating detectors, 124–127  
     geometry of, 119–127  
     massive MIMO, 381–383  
     problem, 119–124  
     subspace detector, 124  
 Signal processing, 159–173  
     centralized process, 256  
     chain, 277–278, 281, 285  
     decentralized process, 256  
     in multistatic radar systems, 256–257  
     POMDP, 172–173  
     sequential hypothesis testing, 168–172  
         anticipation, 173  
         binary, 169–171  
         with multiple hypotheses, 171–172  
         partially observable Markov decision  
             processes, 172–173  
     waveform, 159–168, 328–346  
         constant modulus constraints, 168  
         deterministic, known target impulse response,  
             160–163  
         random target impulse response, 163–168  
         shape, 168  
 Signal Processing at RF (SPAR), 16

- Signal-to-interference-plus-noise ratio (SINR), 160–163, 350–351, 423–424  
 average harvested power *vs.*, 490*f*  
 balancing problem, 481  
 FF technique, 466–467  
 information receivers, 485–487, 490–492  
 large-scale fading factors, 388–389  
 massive MIMO precoding techniques, 377*t*  
 maximization, 425–426
- Signal-to-noise ratio (SNR)  
 average maximum balanced, 471, 472*f*  
 balancing problem, 446  
 input, 599  
 massive MIMO precoding techniques, 377*t*  
 maximization with CSI, 408–410  
 MF output, 328–329  
 output, 599  
 radar, 354–355  
 upper asymptote, 359*f*
- Simultaneous wireless information and power transfer (SWIPT), 479  
 beamforming design, 482–502  
 co-located IRs and ERs, 483*f*, 497–502  
 IRs and ERs separation, 483*f*, 484–491  
 MISO system, 482, 483*f*  
 secret information, 491–497, 496*f*  
 CSI acquisition at transmitter, 503–504  
 energy harvesting, 479, 482, 484  
 multiantenna, 481–482  
 multipoint-to-multipoint, 502–503  
 physical-layer security, 491–497  
 power minimization problem, 481  
 receiver, 479–481, 480*f*  
 WPCN, 503
- Single-carrier equalization  
 postchannel approach, 450–452, 451*f*  
 prechannel approach, 460–461
- Single chip RF system, 318, 318–319*f*
- Single frequency network (SFN) radar, group sparsity, 306–311, 307–308*f*  
 EGOMP method, 309, 312*f*  
 signal model, 308–309  
 verification, 309–311, 310–312*f*
- Single-input multi-output (SIMO) systems, 409  
 group sparsity, 302–304, 303–305*f*
- Single measurement vector (SMV), 555–557
- Single-user networks  
 frequency-flat channels, 406–407, 407*f*  
 partial CSI, 412–414  
 SNR-maximization with perfect CSI, 408–410  
 SNR-per-unit-power maximization, 410–411  
 frequency-selective channels, 423–426
- Singular value decomposition (SVD), 525–527  
 SINR. *See* Signal-to-interference-plus-noise ratio (SINR)
- SLA. *See* Sparse linear array (SLA)
- Slepian formula, 246
- SNR-per-unit-power maximization, 410–411
- Software defined radio (SDR), 256
- SORTE algorithm, 551
- Sparse arrays, 277–278, 291
- Sparse Bayesian learning (SBL), 520, 536
- Sparse iterative covariance-based estimation (SPICE), 529–531  
 generalized least squares, 528–529, 550–551, 553–555  
 second order cone programs, 529–530
- Sparse learning via iterative minimization (SLIM), 519, 527–528
- Sparse linear array (SLA), 551–552  
 covariance fitting criterion, 550–551  
 data model, 538, 542  
 direction-of-arrival estimation, 512–513  
 EMaC, 548  
 GLS optimization problem, 567–568  
*M*-element, 538  
 SDP, 546
- Sparse methods, DOA estimation, 510, 533  
 gridless sparse methods, 510, 537–538  
 ADMM, 570–571  
 ANM *vs.* GLS, 566–568  
 computational issues and solutions, 568–571  
 data model, 538  
 dimensionality reduction technique, 569–570  
 multiple snapshot case, 553–562  
 reweighted atomic norm minimization, 562–566  
 single snapshot case, 541–548, 550–552  
 Vandermonde decomposition of Toeplitz covariance matrices, 537–541
- off-grid sparse methods  
 dynamic grid, 536–537  
 fixed grid, 533–536
- on-grid sparse methods, 510, 522  
 coherence-inhibiting technique, 532  
 convex relaxation, 524–527  
 data model, 522–523  
 dimensionality reduction technique, 525–527  
 GLS, 528–529  
 $l_{2,0}$  optimization, 523–524  
 $l_{2,q}$  optimization, 527–528  
 maximum likelihood estimation, 531–532  
 multiple measurement vectors, 523  
 selection, 532–533

- Sparse methods, DOA estimation (*Continued*)
  - single measurement vectors, 523
  - singular value decomposition, 525–527
  - SPICE, 529–531
- Sparse reconstruction techniques, 278
- group sparsity, 300–311, 301*f*
    - EGOMP method, 309, 312*f*
    - group model, 302
    - MIMO radar network, 300, 305–306, 305–306*f*
    - SFN radar, 306–311, 307–308*f*, 310–312*f*
    - SIMO radar network, 302–304, 303–305*f*
  - spatial sparsity, 291–299
    - DOA, antenna array, 291–293, 292–294*f*
    - 3D-SAR, 294–299, 295*f*, 297*f*, 299–300*f*
  - spectral sparsity, 287–291
    - gapped frequency band, 287–288*f*
    - recovery of missing/corrupted information, 287–289, 287–289*f*
    - SAR image, 289*f*
    - sub-/co-prime sampling, 289–291
    - traditional signal processing, 290*f*
  - temporal sparsity, 280–286
    - Doppler frequency, 283–286, 286*f*
    - matched filter approach, 282*f*
    - sampling in range, 281–286
  - Sparse total least-squares (STLS) approach, 534
  - Spatial aliasing, 591
  - Spatial filtering, 587
  - Spatial modulation (SM), 35–38, 394–395
  - Spatial sparsity, 291–299
    - DOA, antenna array, 291–293, 292–294*f*
    - 3D-SAR, 294–299, 295*f*, 297*f*, 299–300*f*
  - Spectral efficiency (SE), massive MIMO, 368
  - Spectral sparsity, 287–291
    - gapped frequency band, 287–288*f*
    - recovery of missing/corrupted information, 287–289, 287–289*f*
    - SAR image, 289*f*
    - sub-/co-prime sampling, 289–291
    - traditional signal processing, 290*f*
  - Spectrum-shaped FM waveforms, 24–31
  - SPGL1 algorithms, 302
  - Spherical Interpolation (SI) method, 261–262
  - Spherical Intersection (SX) method, 261–262
  - SPICE. *See* Sparse iterative covariance-based estimation (SPICE)
  - Spinor
    - expansion, complex vector as, 94–98
    - interpretation, 97*f*
    - operation of, 106*f*
    - vectors rotation via, 89–92
  - Square-root LASSO (SR-LASSO), 518, 531
  - Steering vector, 114, 115–116*f*, 590
  - Stepped frequency radar, 331–332, 332*f*
    - pseudo-random, 333–339, 334*f*
    - random, 321, 332, 343, 345*f*
  - Stochastic geometry technique, 346–354, 347*f*
    - interference analysis, 349–350
    - interference statistics, 350
    - lattice model, 348–349
    - performance analysis and optimization, 350–354, 351–352*f*, 354*f*
    - Poisson point process model, 347–348, 348*f*
    - ranging success probability, 346, 350–351, 351*f*
    - spatial success probability, 352–353, 352*f*
  - Sub-prime sampling, spectral domain, 289–291
  - Subspace methods, 525, 541, 551
  - Sum-rate maximization, 419–420, 430–431
    - asynchronous networks, 455–457
    - synchronous networks, 438–439
  - Superdirective (SD) beamforming, 600–601, 602*f*
  - SWIPT. *See* Simultaneous wireless information and power transfer (SWIPT)

## T

- Target detection, 257–258
- ANMF detector, 240–242
    - centralized coherent detector, 257*f*
    - centralized noncoherent detector, 258*f*
    - decentralized detector, 258*f*
    - detection performance, 242–245
    - hypothesis testing problem for, 240–245
  - Target localization, 259–262
  - Target resolution, 258–259, 259*f*
  - TDBC technique. *See* Time division broad cast (TDBC) technique
    - Temporal filtering, 587
    - Temporal sparsity, 280–286
      - sparse sampling in range, 281–282
        - and Doppler frequency, 283–286, 286*f*
        - matched filter approach, 282*f*
    - Time-bandwidth product, 328–329
    - Time division broad cast (TDBC) technique, 427, 440
    - Time-division duplexing (TDD) system, 367–368, 375, 384–387, 385*f*
    - Time-indexed model, 187
    - Time-of-arrival (TOA), 198–199, 254–255
    - Time-switching (TS) receiver, SWIPT, 480, 480*f*
    - Toeplitz covariance matrices, Vandermonde decomposition, 537–541
    - Total MSE minimization, two-way network beamforming, 452–455

Total power minimization, two-way network  
 beamforming  
 asynchronous networks, 457–459  
 synchronous networks, 431–435  
 Transmitter distortion, 13–14  
 desired and actual phase transitions, 14/*f*  
 pulse shape after, 14/*f*  
 spectral content, 13/*f*  
 Transmitter effects, 12–15  
 Transmitter-in-the-loop optimization, 31–35  
 Transmitter power, 354–355  
 Traveling wave tubes (TWTs), 5–7  
 model-in-the-loop optimization for, 33, 33–34/*f*  
 Trivector  
 attributes, 77  
 geometric view of, 52–53, 53/*f*  
 2D array, direction-of-arrival, 293, 511–513, 512/*f*  
 Two-way network beamforming, 427  
 asynchronous networks, 440–464  
 end-to-end channel model, 442  
 joint prechannel/postchannel equalization,  
   461–464, 462/*f*  
 max-min SNR fair design approach, 443–450  
 multicarrier equalization, 443  
 single-carrier postchannel equalization,  
   450–452, 451/*f*  
 single-carrier prechannel equalization,  
   460–461  
 sum-rate maximization, 455–457  
 total MSE minimization, 452–455  
 total power minimization, 457–459  
 frequency-selective transceiver-relay links,  
   464–467  
 filter-and-forward relaying, 466–467  
 OFDM-based channel equalization, 465–466  
 geometry, 469/*f*  
 numerical examples, 468–472, 473/*f*  
 OFDM-based, 444/*f*  
 sum-rate curves vs. total transmit power,  
   471–472/*f*  
 synchronous networks, 428–440  
 individual power constraints, 439–440  
 max-min SNR fair design approach, 435–437  
 sum-rate maximization, 438–439  
 TDDBC vs. MABC, 440  
 total power minimization, 431–435  
 with two transceivers, 429/*f*

**U**

Ultralow sidelobe (ULS) waveforms, 24  
 annotated range profile using, 30/*f*  
 autocorrelation, 27/*f*

Range-Doppler ambiguity function, 28/*f*  
 spectral content of, 28/*f*  
 Uniform linear array (ULA), 546–547, 551–552  
 ANM, 545–546  
 direction-of-arrival estimation, 512–513, 513/*f*  
 equivalences/equalities, 551–552  
*M*-element, 538  
 Unit bivector, 90  
 Unoriented real number, 51

**V**

Vandermonde decomposition, Toeplitz covariance  
 matrices, 537–541  
 Vectors, 52  
 addition, 67  
 arbitrary vector, 59–60  
 area of parallelogram, 63/*f*  
 associative product of, 61–66  
 attributes, 73/*f*  
 bivector, 73  
 components of, 54/*f*  
 coordinate-free treatment, 74  
 coplanar vectors, 87/*f*  
 cross product in three dimensions, 67/*f*  
 equal-length vectors, 64/*f*  
 geometric product, 66–74  
 inner product, 69  
 inverse vector, 60  
 magnitude of cross product, 72, 72/*f*  
 multiplication, 57–74  
 multiplicative inverse, 68  
 multivectors, 67  
 nonassociative product of, 57–61  
 orientation, 72  
 orthogonal projection, 86, 87/*f*  
 orthogonal vectors, 70  
 orthonormal vectors, 103  
 outer product, 69  
 parallel vectors, 70  
 relative directions, 62  
 rotation via spinors, 89–92  
*r*-vector, 80  
 sum of, 90  
 three-dimensional space, 59  
 in three dimensions, 64/*f*  
 unit vector, 72–73

**W**

Waveform  
 agility, 18–19  
 design, 10–12, 16–17, 22–24, 159–168

- Waveform (*Continued*)  
target detection with known impulse response, 174–175  
target detection with known nuisance parameter, 179–185  
optimization process, 25  
performance metrics, 9–12  
pseudo-random stepped frequency, 333–339, 334f, 339f  
signal processing, 328–346  
Welch bound, 279  
White noise gain (WNG), 599–601  
White Rabbit network, 263  
Wideband beamforming, 593–597  
approaches in, 597–605
- Lagrange multiplier method, 595–597, 599–600  
LCMV, 601–605  
least squares, 595–596  
superdirective, 600–601, 602f  
Wideband MIMO radar, 38–44  
Wiener filter, 605, 609–610  
Wireless information transfer (WIT), 479–481, 503  
Wireless powered communication network (WPCN), 503  
Wireless power transfer (WPT), 479–482, 503  
Worst case analysis, 524

**Z**

- Zero forcing (ZF) precoding, 377–380  
Ziv-Zakai bound, 198–199

# Academic Press Library in Signal Processing

## Volume 7: Array, Radar and Communications Engineering

This seventh volume in the Academic Press Library in Signal Processing, edited and authored by world-leading experts, gives a review of the principles, methods and techniques of important and emerging research topics and technologies in Array, Radar, and Communications Engineering.

### Features:

- Quick tutorial reviews of important and emerging topics of research in Array, Radar, and Communications Engineering
- Presents core principles in signal processing theory and shows their application
- Comprehensive references to journal articles and other literature on which to build further, more specific and detailed knowledge
- Edited by leading people in the field who, through their reputation, have been able to commission experts to write on a particular topic

### With this reference source you will:

- Quickly grasp a new area of research
- Understand the underlying principles of a topic and its application
- Ascertain how a topic relates to other areas and learn of the research issues yet to be resolved

### Also available in the Academic Press Library in Signal Processing:

Volume 1: Signal Processing Theory and Machine Learning, 9780123965028

Volume 2: Communications and Radar Signal Processing, 9780123965004

Volume 3: Array and Statistical Signal Processing, 9780124115972

Volume 4: Image, Video Processing and Analysis, Hardware, Audio, Acoustic, and Speech Processing, 9780123965011

Volume 5: Image and Video Compression and Multimedia, 9780124201491

Volume 6: Image and Video Processing and Analysis and Computer Vision, 9780128118894



ACADEMIC PRESS

An Imprint of Elsevier  
[elsevier.com/books-and-journals](http://elsevier.com/books-and-journals)

ISBN 978-0-12-811887-0



9 780128 118870