Listening to Distances and Hearing Shapes: Inverse Problems in Room Acoustics and Beyond

THÈSE N^{O} 6623 (2015)

PRÉSENTÉE LE 12 JUIN 2015
À LA FACULTÉ INFORMATIQUE ET COMMUNICATIONS
LABORATOIRE DE COMMUNICATIONS AUDIOVISUELLES
PROGRAMME DOCTORAL EN INFORMATIQUE ET COMMUNICATIONS

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Ivan DOKMANIĆ

acceptée sur proposition du jury:

Prof. E. Telatar, président du jury
Prof. M. Vetterli, directeur de thèse
Prof. H. Bölcskei, rapporteur
Prof. D. Donoho, rapporteur
Prof. M. Unser, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2015

Acknowledgments

> I get by with a little help from my friends.
>
> *The Beatles*

Martin, thank you for giving me the opportunity to write this. The reason I came to EPFL is because I read your paper in SPMag and I just wanted to work with whoever wrote it. I didn't even know that you're at EPFL or that EPFL is in Lausanne. It turns out, coming here was the best decision I ever made. Thank you for the friendly mentoring, for opening my eyes to the beauties of research, for teaching me how to spot interesting problems, and how to ask the right questions. Your talent and drive don't cease to inspire me (nor to remind me that there is always someone who works harder).

It was an honor (and a challenge!) to have Emre Telatar, Michael Unser, Helmut Bölcskei and Dave Donoho on my thesis committee. Thank you for being the first ones to scrutinize these pages!

During my PhD, I had the pleasure to work with two brilliant Frenchmen: Rémi Gribonval and Laurent Daudet. Thank you for giving me the opportunity to learn from you (Rémi, I promise to finish the paper very soon). Yue Lu helped me with my first research steps in the lab, and he has been a steady source of valuable advice ever since. Thank you for that Yue!

This thesis is a coalescence of numerous interactions with friends from LCAV and from around EPFL. Thank you Mihailo and Vojin for friendship, trips, dinners, and intoxicated discussions (and Mihailo for coauthorship!), Reza for friendship and coauthorship, Yann for friendship and sharing your infinite knowledge about matrices, Andreas for your infectious laughter and the perceptual perspective on things, Marta for bringing the music to the lab. Thank you Robin for cobrewing, coauthoring, and masterful Python broadcasting, Jovanče for great memories from numerous skiing weekends, Runwei for all the hot pots at your place. Thanks Andrea and Vincent for that summer in Seattle. What an experience that was!

Thank you Dirk, Benjamin, Loïc, Mitra, Niranjan, Zhou, Yun, Paolo, Andrea, Amina, Adam, Hesam, Jay, Hanjie, Alireza, Tao for being part of this voyage. This goes to all others whom I have unjustly forgotten (forgive me) and who have made my adventure in Lausanne an awesome one.

Then Juri, my office mate: Thanks for a million things. For the numerous *nuits blanches* when Martin would find us incapacitated in the morning trying hard to submit that paper, for

the great fun we had at conferences, for all the discussions we had over the years. I learned a lot from you: How to find interesting problems in unexpected places, how to constantly look outside our field. I learned a thing or two about business, food, and clothes (so that I spend much more now, but I still can't dress). Sharing the office with someone so good at so many things motivated me to work harder myself. Best of luck!

Thanks Jacqueline for saving my rear end $(n+1)$ times. The stoicism with which you handled the chaos I was leaving behind is extraordinary.

Thank you Ivo, Krešo Zubalo, Ervin, Zvone, Mali, Edo, Mislav, Goran: *Karavana.* Frustratingly, my best summer happened when I decided to go to Switzerland. The legendary Caravan—a happy, optimistic, life-loving, singing, guitar-playing crowd that ruled the Adriatic. But now the spirit of the Caravan is freed, and nothing can stop it! You can all hear the siren song, I know. Special thanks to Ervin for letting me live in his house when I was back in Zagreb. *Hvala prika!*

Thanks to Bebo and Lucia! What neighbors and friends to have... great fun, food and wine! I hope we keep this going. Thank you Sanja for never, not once being in a bad mood since we met.

I dedicate this thesis to Metka, my beloved, beautiful girl. You made my life beautiful too! We have the big things all figured out, now for the small ones. The world is ours, of course, and knowing that I have *you* makes me peaceful. *Tuturutu!*

All my achievements, and in particular this thesis, begin and end with my family. Thank you mom and dad for your unconditional love and support, but above all for the absolute, liberating trust you put in me. Thanks Marko for being a great brother and friend, for partying, laying down the groove, being a good sport, and for always recovering and forgiving, no matter how bad we fought.

Abstract

A central theme of this thesis is using echoes to achieve useful, interesting, and sometimes surprising results. One should have no doubts about the echoes' constructive potential; it is, after all, demonstrated masterfully by Nature. Just think about the bat's intriguing ability to navigate in unknown spaces and hunt for insects by listening to echoes of its calls, or about similar (albeit less well-known) abilities of toothed whales, some birds, shrews, and ultimately people.

We show that, perhaps contrary to conventional wisdom, multipath propagation resulting from echoes is our friend. When we think about it the right way, it reveals essential geometric information about the sources–channel–receivers system. The key idea is to think of echoes as being more than just delayed and attenuated peaks in 1D impulse responses; they are actually additional sources with their corresponding 3D locations. This transformation allows us to forget about the abstract *room*, and to replace it by more familiar *point sets*. We can then engage the powerful machinery of Euclidean distance geometry. A problem that always arises is that we do not know *a priori* the matching between the peaks and the points in space, and solving the inverse problem is achieved by *echo sorting*—a tool we developed for learning correct labelings of echoes. This has applications beyond acoustics, whenever one deals with waves and reflections, or more generally, time-of-flight measurements.

Equipped with this perspective, we first address the "Can one hear the shape of a room?" question, and we answer it with a qualified "yes". Even a single impulse response uniquely describes a convex polyhedral room, whereas a more practical algorithm to reconstruct the room's geometry uses only first-order echoes and a few microphones.

Next, we show how different problems of localization benefit from echoes. The first one is multiple indoor sound source localization. Assuming the room is known, we show that discretizing the Helmholtz equation yields a system of sparse reconstruction problems linked by the common sparsity pattern. By exploiting the full bandwidth of the sources, we show that it is possible to localize multiple unknown sound sources using only a single microphone. We then look at indoor localization with known pulses from the geometric echo perspective introduced previously. Echo sorting enables localization in non-convex rooms without a line-of-sight path, and localization with a single omni-directional sensor, which is impossible without echoes.

A closely related problem is microphone position calibration; we show that echoes can help even without assuming that the room is known. Using echoes, we can localize arbitrary numbers of microphones at unknown locations in an unknown room using only one source at an unknown location—for example a finger snap—and get the room's geometry as a byproduct.

Our study of source localization outgrew the initial form factor when we looked at source localization with spherical microphone arrays. Spherical signals appear well beyond spherical microphone arrays; for example, any signal defined on Earth's surface lives on a sphere. This resulted in the first slight departure from the main theme: We develop the theory and algorithms for sampling sparse signals on the sphere using finite rate-of-innovation principles and apply it to various signal processing problems on the sphere.

One way our brain uses echoes to improve speech communication is by integrating them with the direct path to increase the effective useful signal power. We take inspiration from this human ability, and propose acoustic rake receivers (ARRs) for speech—microphone beamformers that listen to echoes. We show that by beamforming towards echoes, ARRs improve not only the signal-to-interference-and-noise ratio (SINR), but also the perceptual evaluation of speech quality (PESQ).

The final chapter is motivated by yet another localization problem, this time a tomographic inversion that must be performed extremely fast on computation- and storage-constrained hardware. We initially proposed the *sparse pseudoinverse* as a solution, and this led us to the second slight departure from the main theme: an investigation of the properties of various norm-minimizing generalized inverses. We categorize matrix norms according to whether or not their minimization yields the MPP, and show that norm-minimizing generalized inverses have interesting properties. For example, the worst-case and average-case ℓ^p-norm blowup is minimized by generalized inverses minimizing certain induced and mixed matrix norms; we call this a *poor man's* ℓ^p *minimization.*

Keywords: Echoes, Euclidean distance geometry, Euclidean distances matrices, multipath, room impulse response, inverse problems, acoustics, room geometry reconstruction, sound source localization, indoor localization, sparsity, sphere, finite rate of innovation, sampling, shot noise, acoustic rake receiver, perceptual evaluation of speech quality (PESQ), beamforming, Moore-Penrose pseudoinverse, alternative generalized inverse, alternative dual frame, sparse pseudoinverse

Résumé

Un thème central de cette thèse est l'utilisation d'échos pour obtenir des résultats utiles, intéressants, et parfois surprenants. Le potentiel constructif des échos ne devrait surprendre personne; la Nature, aprè tout, les utilise avec maîtrise. Il suffit de penser aux chauves-souris, dont l'intrigante capacité à se déplacer dans des environnements inconnus et chasser des insectes repose uniquement sur l'écoute de l'écho de ses cris, ou aux exemples similaires (même si moins bien connus) des cétacés, de certains oiseaux, musaraignes, et, finalement, d'êtres humains.

Nous montrons que la présence de multiples échos est, peut-être contrairement aux idées reçues, notre ami. En y pensant de la bonne manière, il révèle des informations géométriques essentielles au sujet des systèmes sources-canaux-récepteurs. L'idée clé est de considérer les échos comme étant plus que des pics retardés et atténués dans des réponses impulsionnelles unidimensionnelles; ils sont en fait des sources additionnelles, avec leurs positions tridimensionnelles correspondantes. Cette transformation nous permet d'oublier la pièce abstraite, et de la replacer par un ensemble de points plus familier. Nous pouvons ensuite mettre en route la puissante machinerie de la géométrie des distances Euclidiennes. Un problème qui survient toujours est que l'on ne connaît pas à priori la correspondance entre les pics et les points dans l'espace. La solution de ce problème inverse est obtenue grâce au tri des échos—un outil que nous avons développé pour apprendre à identifier les échos correctement. Ceci a des applications au-delà de l'acoustique, dès que l'on a affaire à des ondes et des reflections, ou, plus généralement, des mesures du temps de vol.

Equipés de cette perspective, nous abordons d'abord la question "Peut-on entendre la forme d'une pièce?", et y répondons par un "oui" qualifié. Même une seule réponse impulsionnelle décrit une pièce polyédrique convexe de manière unique, alors qu'un algorithme plus pratique pour reconstruire la géométrie de la pièce n'utilise que les échos de premier ordre et quelques microphones.

Ensuite, nous montrons comment différents problèmes de localisation bénéficient des échos. Le premier est la localisation de multiples sources sonores intérieures. Supposant que la pièce est connue, nous montrons que la discrétisation des équations de Helmholtz donne un système de problèmes de reconstruction parcimonieux, liés par une structure parcimonieuse commune. En exploitant toute la largeur de bande des sources, nous montrons qu'il est possible de localiser plusieurs sources sonores inconnues en n'utilisant qu'un seul microphone. Nous nous intéressons ensuite à la localisation intérieure à partir d'impulsions connues, vu sous la perspective des échos géométriques introduite précédemment. Le tri des échos permet la localisation dans une pièce non-convexe et sans chemin en ligne de vue, et la localisation avec un seul senseur omnidirectionnel, ce qui est impossible sans échos.

Un problème intimement lié est la calibration du positionnement de microphones; nous montrons que les échos peuvent aider, même sans supposer que la forme de la pièce est connue. Grâce aux échos, nous pouvons localiser un nombre arbitraire de microphones, à des positions inconnues, en utilisant une seule source elle-même à une position inconnue—par exemple un claquement de doigts—et obtenir la géométrie de la pièce comme résultat annexe. Notre étude de la localisation des sources a dépassé son cadre initial lorsque nous nous sommes intéressés à la localisation d'une source avec un ensemble sphérique de microphones. Les signaux sphériques apparaissent bien au-delà des ensembles sphériques de microphones; par exemple, tout signal défini sur la surface de la Terre réside sur une sphère. En marge du thème principale de la thèse, nous avons développé la théorie et les algorithmes pour échantillonner des signaux parcimonieux sur la sphère, en utilisant les principes du taux d'innovation fini, et les avons appliqués à divers problèmes de traitement du signal sur la sphère.

Une manière dont notre cerveau utilise les échos pour améliorer les communications orales est en les intégrant avec le chemin direct pour augmenter la puissance du signal effectif. Nous nous inspirons de cette capacité humaine, et proposons les récepteurs acoustiques en râteau (RARs) pour la parole—des microphones formant des faisceaux qui écoutent les échos. Nous montrons qu'en formant des faisceaux vers les échos, les ARRs améliorent non seulement le rapport signal sur bruit plus interférence, mais aussi la qualité vocale perçue.

Le chapitre final a été motivé par un autre problème de localisation—cette fois une inversion tomographique qui doit être effectuée extrêmement rapidement sur du matériel à la puissance et au stockage limités. Nous avons proposé comme solution la pseudo-inverse parcimonieuse et ceci nous a mené à notre second léger écart du thème principal: une étude des propriétés de différentes inverses généralisées minimisant une norme particulière. Nous catégorisons les normes matricielles selon si leur minimisation retourne la pseudo-inverse de Moore-Penrose ou non, et montrons que les inverses généralisées minimisant une norme ont des propriétés intéressantes. Par exemple, dans le pire cas et en moyenne, l'augmentation de la norme ℓ^p est minimisée par les inverses généralisées minimisant certaines normes matricielles induites et mixées; nous appelons cette méthode de la minimisation *la minimisation ℓ^p du pauvre.*

Mots-clès: Echos, géométrie des distances Euclidiennes, matrices des distances Euclidiennes, trajets multiples, réponse impulsionnelle d'une chambre, problèmes inverses, acoustique, reconstruction de la forme d'une pièce, localisation des sources sonores, localisation à l'intérieure, parcimonie, sphère, finite rate of innovation, échantillonnage, shot noise, récepteurs acoustiques en râteau, "perceptual evaluation of speech quality (PESQ)", formation de faisceau, pseudo-inverse de Moore-Penrose, inverses généralisées alternatives, "alternative dual frame", pseudo-inverse parcimonieuse

Contents

Acknowledgments **ii**

Abstract **v**

Résumé **vii**

1 Introduction **1**
- 1.1 Inverse Problems in Room Acoustics and Beyond . . . 2
  - 1.1.1 Inverse Problems . . . 2
  - 1.1.2 Room Acoustics . . . 3
  - 1.1.3 Beyond . . . 4
- 1.2 Thesis Outline and Main Contributions . . . 4

2 On Echoes and Distances **11**
- 2.1 Undulating Particles: The Wave Mechanics of Sound and Radio . . . 11
  - 2.1.1 Mass on a Spring . . . 11
  - 2.1.2 Vibrating String . . . 14
  - 2.1.3 Wave Equation of Sound . . . 16
  - 2.1.4 Plane-wave Solution . . . 19
- 2.2 Origin of Echoes . . . 20
  - 2.2.1 Boundary Conditions in Wave Equations . . . 20
  - 2.2.2 Reflections of Ropes . . . 20
  - 2.2.3 Reflections of Sound . . . 22
- 2.3 Room Acoustics and Echo Modeling . . . 24
  - 2.3.1 Green's Function . . . 24
  - 2.3.2 Image Source Method . . . 26
  - 2.3.3 Need for Geometrical Acoustics . . . 29
  - 2.3.4 Image Source Model in Rooms . . . 29
- 2.4 Distances and Euclidean Distance Matrices: From Points to EDMs and Back . . . 30
  - 2.4.1 Essential Uniqueness . . . 31
  - 2.4.2 Reconstructing the Point Set From Distances . . . 33
  - 2.4.3 Orthogonal Procrustes Problem . . . 35
  - 2.4.4 Counting the Degrees of Freedom . . . 35
- 2.5 EDMs as a Practical Tool . . . 36
  - 2.5.1 Exploiting the Rank Property . . . 37
  - 2.5.2 Multidimensional Scaling . . . 39

2.5.3 Semidefinite Programming 40
2.5.4 Performance Comparison of Algorithms 43
2.6 Conclusion 45

3 Can One Hear the Shape of a Room? 47
3.1 Introduction 47
3.2 Can One Hear the Shape of a Drum? 49
3.2.1 Historical Notes and Kac's Paper 51
3.2.2 Counterexamples: Isospectral Drums 52
3.3 Unlabeled Distances 53
3.4 Modeling 55
3.5 Hearing the Shape of a Room with One Microphone 56
3.5.1 Characterization of RIRs from Image Source Patterns 57
3.6 Room Geometry Estimation 58
3.6.1 The Shape of a Polygonal Room Using Matrix Analysis 58
3.6.2 Uniqueness of RIR: An Assignment Problem 59
3.6.3 Indoor Localization as a Byproduct 61
3.6.4 A Generational Issue: An Unexpected EDM 61
3.7 Numerical simulations 63
3.8 Hearing the Room With More Than One Microphone 64
3.8.1 Modeling 64
3.8.2 Labeling the Echoes 65
3.8.3 EDM-based approach 65
3.8.4 Subspace-based approach 67
3.8.5 Uniqueness 68
3.8.6 Practical Algorithm 69
3.8.7 Computational Complexity 71
3.8.8 Experiments 72
3.9 Conclusion 74
3.A Proof of Theorem 3.3 76
3.B Room Reconstruction Procedure 77

4 Localization and Position Calibration 79
4.1 Introduction 79
4.2 Sound Source Localization with Finite Elements and Wideband Diversity 81
4.2.1 Problem Setup 82
4.2.2 Discretizing the Wave Equation with the Finite Element Method 82
4.2.3 Sparsity and Wideband Diversity 84
4.2.4 Using Only One Microphone 87
4.2.5 Numerical Experiments 87
4.3 Distance-Based Approaches (or a Tale of Two Applications of Echo Sorting) 89
4.3.1 Multilateration 90
4.3.2 Single-Channel Localization 91
4.3.3 Echo Labeling and Almost Sure Uniqueness 92
4.3.4 Experimental Example 93
4.3.5 The Non-Convex Case: Inverse Method of Images 94
4.3.6 Reflecting Localized Sources 95

4.3.7 Estimating the Source Position . . . 95
4.3.8 Optimizing the Position Estimate . . . 96
4.3.9 Tracking . . . 97
4.3.10 Numerical Simulations for Non-Convex Localization . . . 97
4.4 Position Calibration of Microphone Arrays . . . 98
4.5 Microphone Calibration and Multidimensional Unfolding . . . 100
4.5.1 Unfolding as Matrix Completion . . . 101
4.5.2 Integrating Prior Knowledge . . . 103
4.5.3 Numerical Experiments . . . 103
4.6 How to Localize Ten Microphones in a Finger Snap (With a Little Help From Echoes) 104
4.6.1 Anechoic Calibration . . . 104
4.6.2 Zero-Knowledge Indoor Calibration . . . 105
4.6.3 Minimal Infrastructure for Calibration . . . 106
4.6.4 Algorithm . . . 106
4.6.5 Numerical Experiment . . . 108
4.A Proof of Proposition 4.1 . . . 110

5 Sampling Sparse Signals on the Sphere 111
5.1 Introduction . . . 111
5.1.1 Prior Art . . . 112
5.1.2 Outline and Main Contributions . . . 113
5.2 Harmonic Analysis on the Sphere and Problem Formulation . . . 113
5.2.1 Spherical Harmonics . . . 113
5.2.2 Rotations and Convolutions on the Sphere . . . 114
5.2.3 Problem Formulation . . . 116
5.3 Sampling Spherical FRI Signals . . . 116
5.3.1 From Samples to the Fourier Transform . . . 117
5.3.2 The Data Matrix . . . 118
5.3.3 Reconstruction by Generalized Annihilating Filtering . . . 119
5.3.4 Sampling Efficiency and Relation to Prior Work . . . 122
5.3.5 Denoising Strategies . . . 123
5.3.6 Cramér-Rao Lower Bound . . . 124
5.4 Applications . . . 124
5.4.1 Sampling Diffusion Processes on the Sphere . . . 125
5.4.2 Shot Noise Removal . . . 127
5.4.3 Sound Source Localization . . . 130
5.5 Conclusion . . . 132
5.A Annihilating Property . . . 134
5.B Computation of the Cramér-Rao Lower Bound . . . 134
5.C Rank of the annihilating matrix . . . 135

6 Acoustic Rake Receiver 139
6.1 Introduction . . . 139
6.1.1 Wireless vs Speech . . . 139
6.1.2 Related Work . . . 141
6.1.3 Main Contributions and Limitations . . . 141
6.2 Signal model . . . 142

6.3 Beamforming Preliminaries . . . 145
6.3.1 Delay-and-Sum Beamformer . . . 146
6.3.2 Maximum Signal-to-Interference-and-Noise Ratio Beamformer . . . 146
6.4 Acoustic Rake Receivers . . . 147
6.4.1 Delay-and-Sum Raking . . . 147
6.4.2 One-Forcing Raking . . . 148
6.4.3 Max-SINR Raking . . . 148
6.4.4 Max-UDR Raking . . . 149
6.4.5 SINR Gain from Raking . . . 150
6.5 Comments on Finding and Tracking the Echoes . . . 151
6.5.1 Known Room Geometry . . . 152
6.5.2 Acoustic Image Source Localization . . . 152
6.6 Numerical Experiments . . . 153
6.6.1 Simulation Setup . . . 153
6.6.2 Results . . . 154
6.7 Time-Domain Formulation . . . 157
6.7.1 Signal Model . . . 158
6.7.2 Time-Domain Rake Beamformers . . . 159
6.7.3 Numerical Results for TD Beamformers . . . 161
6.8 Summary and Conclusion . . . 162
6.A Theorem 6.1 . . . 164

7 Norm-Minimizing Generalized Inverses **165**
7.1 Introduction . . . 165
7.1.1 Prior Art . . . 167
7.1.2 Contributions and the Chapter Outline . . . 168
7.2 Definitions and Known Results (Preliminaries) . . . 169
7.2.1 Generalized Inverses . . . 169
7.2.2 Correspondence Between Generalized Inverses and Dual Frames . . . 169
7.2.3 Characterization with the Singular Value Decomposition (SVD) . . . 170
7.2.4 The Moore-Penrose Pseudoinverse (MPP) . . . 170
7.2.5 Generalized Inverses Minimizing Matrix Norms . . . 172
7.3 Preliminary Results . . . 174
7.3.1 Summary and Visualization of Matrix Norms . . . 175
7.4 Classes of Norms Yielding the Moore-Penrose Pseudoinverse . . . 175
7.4.1 Unitarily invariant norms . . . 177
7.4.2 Left unitarily invariant norms . . . 178
7.4.3 Left unitarily invariant norms on the product operator . . . 179
7.4.4 Classical norms leading to the MPP . . . 179
7.5 Norms that Almost Never Yield the MPP . . . 180
7.5.1 Sparse Pseudoinverse . . . 181
7.5.2 Unbiasedness of Generalized Inverses . . . 184
7.5.3 Poor man's ℓ^p minimization revisited . . . 189
7.6 Matrices Having the Same Inverse for Many Norms . . . 190
7.6.1 Matrices which MPP has Entries on the Unit Circle . . . 190
7.6.2 Matrices with a highly sparse generalized inverse . . . 193
7.7 Computation of Norm-Minimizing Generalized Inverses . . . 193

7.7.1 Norms that Reduce to the Vector Case 194
7.7.2 ADMM for Norms that do not (Easily) Reduce to the Vector Case 194
7.7.3 Alternating Direction Method of Multipliers 194
7.7.4 Generic ADMM for matrix norm minimization 196
7.8 Conclusion 197
7.A Proof of Corollary 7.2 198

Conclusion **201**

Bibliography **205**

Curriculum Vitae **218**

Chapter 1

Introduction

> There's nothing you can do that can't be done.
>
> *The Beatles*

How does one hear a shape? Ask a bat: they learn about their surroundings by listening to echoes. People are fascinated by the bat's ability—just think about the pop culture icons like Daredevil and Batman. In the 2008 Hollywood blockbuster *Batman: The Dark Knight*, Batman and his accomplices use cellphones of unsuspecting citizens to *see* the spaces on the other side of the line.

This is far fetched, but not *that* far. One of the central questions in this thesis is "Can one hear the shape of a room?", and we show that what happens in *The Dark Knight* may have crossed the line between fantasy and reality. The trick is to listen to distances.

So how does one listen to a distance? Here is one way: when I was a kid, my dad taught me to figure out how far the lightning had struck by counting the seconds between the lightning and the thunder. Another way is to listen for the echoes of your footsteps from a nearby wall; if you ever did that, you must have noticed how the gap between the footstep and its echo closes as you approach the wall.

In both these examples, we convert *time* to *distance.* We can do it because we know the speed of sound. It should not seem impossible that if we replaced ears and back-of-the-envelope computations with microphones connected to a computer and sophisticated algorithms, we could answer much more complicated questions than how far the lightning had struck. We could for example reconstruct shapes of objects or rooms from echoes, very much like in the *The Dark Knight*, or we could navigate a flying robot by sound only.

We listen to distances by listening to echoes, and it allows us to reason about the geometry of our environment (just like the bats do!). The good news is that there are plenty of echoes to go around. We spend a lot of time in rooms, and rooms serve many purposes; but one "purpose" often goes unnoticed: a well designed room makes having a conversation easier. This is because in rooms we have echoes, and our brain can integrate the early echoes with the direct sound, thus increasing the effective signal-to-noise ratio (in fact, we will show that computers can do very much the same). This thesis is about taking inspiration from people, bats and the Dark Knight, to harness the potential of echoes in solving interesting problems.

1.1 Inverse Problems in Room Acoustics and Beyond

The kind of problems we address is indicated in the subtitle of the thesis. There are three parts in the phrase “Inverse problems in room acoustics and beyond” that deserve an explanation: *inverse problems*, *room acoustics*, and *beyond*.

1.1.1 Inverse Problems

A different name for inverse problems could be simply *problems*. The qualifier *inverse* is added to point out that there is a related problem, perhaps better known and more thoroughly studied, to which the one that we study is somehow opposite. This other problem is then called the *forward* problem. Sometimes, it is arbitrary which is called forward and which is called inverse. Often what we measure is from the forward problem, and we try to invert it to learn something about the mechanism that produced the measurements.

A good rule of thumb for designating problems as forward or inverse is that the forward role is typically played by a physical process (more generally, by Nature). The Big Bang left behind a trail of ripples in the background radiation of the Universe. To go from the big bang to the ripples was Nature’s job. To solve the forward problem would then mean to devise the computational machinery that can simulate the measurements of any parameter of interest, given the conditions of the big bang (this one is conceptually simple and, for the time being, practically impossible). The inverse problem could be learning about the parameters of the Universe right after the big bang from the measurement of these residual cosmic ripples [2].

To get a better idea of what we mean by inverse problems, consider the following list of forward-inverse problem pairs:

(i) Given Earth’s gravity acceleration g, and the initial speed of a projectile, compute its trajectory; given the measurements of the projectile’s position at k distinct (but known) times, compute g (to make it harder, change *known* to *unknown*);

(ii) Given the room geometry and the properties of walls, as well as the source or sound, compute the sound field for all times at all points in the region; given the measurements of the sound pressure at three points, compute the location of the sound source (with or without knowing the boundary conditions);

(iii) Given the geometry of the ultrasonic transducer array placed around a female breast, and the density distribution of the breast tissue, compute the waveforms transducers record when each of them fires a pulse; given the array geometry and the waveforms, compute the density distribution in the breast;

(iv) Given the shape of a drumhead (a membrane), compute its resonant frequencies; given the set of resonant frequencies of a drum, find its shape;

(v) Given details about a guitar-amplifier setup and geometry, compute the feedback waveform; given the studio recording of “I feel fine” by The Beatles, reconstruct the distance of Lennon’s Gibson J-160E from the amplifier’s speaker over time.[1]

We typically think of an inverse problem as being more difficult than the forward one, but this is not necessarily the case. For example, the forward problem could be: Given the coefficients of a univariate polynomial, find its roots. Then the inverse problem—given the roots, find the polynomial—is much simpler.

Although it may seem banal, this example illustrates the concept of *well-posedness*—one of the major concepts related to inverse problems. Without further qualification, the problem of finding the polynomial given its roots is not well-posed, because it does not have a unique solution. It is easy to find *a* solution; but if we have a particular polynomial in mind that generates the given set of roots, it is hopeless to expect that we will find it, as any rescaling of the coefficients leaves the roots intact.

The concept of well-posedness was formalized by Hadamard [89] who called *ill-posed* any problem that fails to meet any of the following requirements:

- Existence of a solution,
- Uniqueness,
- Stability with respect to data,

Hadamard had strong reservations about ill-posed problems (he thought that they bear no physical significance). But the field has progressed tremendously, and nowadays most (if not all) interesting inverse problems are, in fact, ill-posed.

1.1.2 Room Acoustics

Room acoustics is a sub-field of acoustics that deals with how the sound spreads in rooms. In itself, it addresses two very different questions. The first one is about the physics: What are the wave mechanics of sound generation and propagation in enclosures? This includes studying specular and diffuse reflections, edge diffraction, scattering by various structures, and influence of a range of other parameters.

The second one is perceptual: How pleasing do we find the acoustics of a particular space? It could also be about how conducive acoustics of a space are to specific purposes: listening to a concert of classical music, listening to a jazz concert, listening to a lecture, or perhaps just chatting in a café. It uses what we know about sound propagation from the first question and combines this knowledge with psychoacoustics and ergonomy.

In this thesis, we are mainly interested in the physical aspect of the room-acoustical universe. We aim to use the properties of wave propagation to solve interesting inverse problems. We restrict ourselves even further to deal only with early reflections. The models that we use are models that are very common in room acoustics, but that have little to do with perception (as my colleagues from the hi-fi end of the spectrum rarely fail to observe).

---

[1] For an idea, take a look at `http://nbviewer.ipython.org/github/LCAV/SignalsOfTheDay/blob/master/I_feel_fine/I%20feel%20fine.ipynb`, created by Paolo Prandoni.

1.1.3 Beyond

So far we talked about echoes of sound in rooms. But what we really exploit is the distance geometry of reflections which is common for all wave phenomena. In particular, our results are applicable to multipath propagation of radio waves.

The geometric framework that we develop is useful whenever we deal with sets of distances without labels. Straightforward examples are in sensor network calibration, but an exciting insight is that the same kind of problems appear in crystallography.

Two chapters of the thesis deal with inverse problems that depart from the echo-distance framework. The first one grew out of thinking about sound source localization with spherical microphone arrays. That led us to the development of spherical finite-rate-of-innovation sampling theory for sampling sparse signals on the sphere. The second part grew out of thinking about a tomographic approach to touch displays. We proposed the *sparse pseudoinverse* to deal with very limited computation and storage budget on embedded hardware. This led us to a deeper study of norm-minimizing generalized inverses.

1.2 Thesis Outline and Main Contributions

Chapter 2: On Echoes and Distances To make the thesis complete and self-contained, we start from the basic physical principles. In the first part of Chapter 2, we explain how echoes are created by waves and how we model them. This sets the stage for the introduction of a key tool: the image source model. We introduce it first as an exact tool to solve PDEs on a few particular domains, and then as a tool in geometric acoustics that can handle any geometry. The second part of the chapter talks about Euclidean distance geometry (EDG); in fact, it mostly deals with a particular tool that we find useful over and over: Euclidean distance matrices (EDM). In the remaining chapters we use the conjunction of the image source model and EDG as a powerful device for extracting geometric information in rooms. Chapter 2 can serve either as a reminder, or as a primer on these matters for an uninitiated reader.

Summary of Contributions in Chapter 2

- It is certainly unusual to claim contributions in a chapter that reviews the background material, but we would like to point one out: the perspective on EDM completion by counting the degrees of freedom seems novel.

Chapter 3: Can One Hear the Shape of a Room? In 1966, Mark Kac asked “Can one hear the shape of drum?”. What he actually wanted to elucidate is whether the problem of computing a drum’s shape from its resonant frequencies is well-posed.

The central part of this thesis treats a transposition of Kac’s question to rooms. Imagine that you are blindfolded inside an unknown room; you snap your fingers and listen to the room’s response. Can you hear the shape of the room? Some people can do it naturally, but can we design computer algorithms that hear rooms? Unlike Kac, when we ask “Can one hear the shape of a room?”, we want to know if one could find the shape of a room from room impulse responses (RIRs), not from its resonant frequencies. But instead of approaching the question by tools from functional analysis, we argue that a more useful and practical perspective is through echoes and image sources, with a strong flavor of geometrical acoustics.

We first present an algorithm that reconstructs the geometry of a convex polyhedral room from a single room impulse response, that is, using a single microphone. We show that the sets of first- and second-order echoes uniquely describe the room's geometry. Our algorithm relies on strong geometric connections between the first two generations of image sources, and between the corresponding echo arrival times. As we only use a single microphone, we require more than just the first-order echoes. Single-microphone room reconstruction is a surprising result, and a proper time-domain dual to the question of Kac.

Obtaining higher-order echoes from RIRs is challenging. That is why our second algorithm uses only first-order echoes and a few microphones to compute the shape of a room. Furthermore, we show that under mild conditions these first-order echoes provide a unique description of convex polyhedral rooms. The key step is *echo sorting*: assigning echoes to correct walls. In contrast to earlier methods, the proposed algorithm reconstructs the full three-dimensional geometry of the room from a single sound emission, and with an arbitrary geometry of the microphone array; as long as the microphones can hear the echoes, we can position them as we want. Besides answering a basic question about the inverse problem of room acoustics, our results find applications in areas such as architectural acoustics, indoor localization, virtual reality and audio forensics.

Summary of Contributions in Chapter 3

- We design an algorithm that can reconstruct the geometry of a convex polyhedral room from a single RIR. We show that the room is uniquely described by the sets of first- and second-order echoes.
- We design an algorithm that reconstructs the geometry of a convex polyhedral room using only first-order echoes recorded by four or more microphones and prove the almost-sure uniqueness of this description. The algorithm has been verified experimentally in a classroom at EPFL and in an alcove of the Lausanne cathedral.
- In the process, we introduce *echo sorting*—an algorithm that finds the correct echo assignment, useful beyond hearing room shapes.

Chapter 4: Localization and Calibration Source localization is a classic inverse problem. Many methods work well in free space, but are challenged by echoes in rooms. We propose a framework in which the room actually *helps*.

Before addressing echoes geometrically, we first consider them only implicitly: we approach the multiple indoor source localization problem by discretizing the Helmholtz equation. We make an assumption that the room is known and then show how a particular discretization scheme—the finite element method (FEM)—simultaneously solves the PDE and provides us with a sparsifying dictionary, so that the source localization can be addressed using sparse recovery methods. The second important ingredient is what we call the *wideband advantage*. One Helmholtz equation models what happens at one frequency, but for wideband sources we can write the Helmholtz equation at many frequencies. The key is that the sparsity pattern in the source vector remains constant over frequencies. Perhaps surprisingly, this observation enables us to localize multiple sources with arbitrary wideband spectra using only a single microphone.

We continue by showing how echo sorting introduced in Chapter 3 helps extract information in a *known* room. We apply it to indoor localization using only a single omnidirectional sensor (which is impossible in free space), and how to localize in non-convex spaces. We assume that the source emits a pulse whose time of arrival can be measured at the receiver side.

Performing tasks with microphones usually requires knowing their positions. In other words, we need to know the geometry of the microphone array. A compelling method to calibrate the array geometry is to use sources at unknown locations. Interestingly, it is possible to reconstruct the locations of both the sources and the receivers if their number is larger than some prescribed minimum; this was recognized recently in a number of works. The problem itself is an instance of *multidimensional unfolding*, and we propose to solve it using EDM completion, thus obtaining a formulation that, unlike earlier works, easily handles missing distances and various prior information about the array.

Next, we show that in rooms we can considerably reduce the number of sources required for array calibration *even when the room is unknown*; the key observation is that echoes correspond to virtual sources that we get "for free". This enables endeavors such as calibrating the array using only a single source, for example a finger snap. With our technique we can also compute the absolute position of the microphone array in the room. This is in contrast to only knowing it up to a rigid transformation or reflection, which is ensured by other methods. What is more, as a byproduct we get the room's geometry!

Summary of Contributions in Chapter 4

- We show how localization of multiple wideband sound sources in a known room can be framed as a system of sparse reconstruction problems with different system matrices and a common sparsity pattern, and propose algorithms to solve it.
- We show that the *room helps* and that *bandwidth helps*; in a room we can localize more sources than in free space. We show through numerical experimentation that the convex relaxation allows us to localize multiple sources with only a single microphone.
- We apply echo sorting to two indoor localization problems from TOA measurements: 1) indoor localization using a single sensor, 2) indoor localization in non-convex rooms. We propose to solve the second one using the *inverse method of images.*
- We exhibit a microphone position calibration method that can integrate prior uncertain knowledge about the array geometry and handle missing pairwise distances. Next, we address the *zero-knowledge calibration* problem. Consider an *unknown* microphone array in an *unknown room.* We show that the geometry of both can be learned from a single finger snap at an *unknown* location.

Chapter 5: Sampling Sparse Signals on the Sphere Answering questions about sound source localization with spherical microphone arrays spilled over into a comprehensive theory for sampling sparse collections of spikes on the sphere, which we present in Chapter 5.

We develop a sampling theorem and corresponding algorithms that can perfectly reconstruct a collection of spikes on the sphere from samples of their lowpass-filtered observations. Central to our algorithm is a generalization of the annihilating filter method, a tool widely used in array signal processing and finite-rate-of-innovation (FRI) sampling. The proposed algorithm can reconstruct K spikes from $(K + \sqrt{K})^2$ spherical samples; this sampling requirement improves over previously known FRI sampling schemes on the sphere by a factor of four for large K.

We showcase the versatility of the proposed algorithm by applying it to three different problems: 1) sampling diffusion processes induced by localized sources on the sphere, 2) shot noise removal, and 3) sound source localization (SSL) by a spherical microphone array.

In the shot noise removal application, some of the sensors give corrupted readings of the underlying bandlimited function; we show theoretically and experimentally that our sparse sampling algorithm can be applied to detect and correct the malfunctioning sensors. In the SSL application we show how Green's functions induced on the sphere by point sources can be approximated as having a constant shape over source positions. We then use deconvolution to reduce this scenario to a collection of angular spikes.

Summary of Contributions in Chapter 5

- We develop the theory and algorithms for sampling collections of spikes on the sphere, requiring one-fourth the number of samples required by the best previous FRI sampling scheme on the sphere.
- Our algorithms can be applied to remove shot noise from the measurements on the sphere (*i.e.*, to detect malfunctioning sensors). We develop a shot noise removal algorithm based on FRI sampling, and compute exact bounds on the number of detectable malfunctioning sensors.
- We demonstrate in detail how our sampling theory can be used to localize sound sources with spherical microphone arrays.

Chapter 6: Raking the Cocktail Party This chapter is about imitating how people use echoes. It is well established that the echoes improve speech intelligibility [24, 25, 127]. In fact, adding energy in the form of early echoes (approximately within the first 50 ms of the room impulse response (RIR)) is equivalent to adding the same energy to the direct sound [25]. This observation suggests new designs for indoor beamformers, with different choices of performance measures and reference signals.

We present the concept of an acoustic rake receiver—a microphone beamformer that uses echoes to improve the noise and interference suppression. The rake idea is well-known in wireless communications; it involves constructively combining different multipath components that arrive at the receiver antennas. Unlike spread-spectrum signals used in wireless communications, speech signals are not (near-)orthogonal to their shifts or to other speech signals. Therefore, we focus on the spatial, rather than temporal structure. Instead of explicitly estimating the channel, we create correspondences between early echoes in time and image sources in space. These multiple sources of the desired and the interfering signal offer additional spatial diversity that we can exploit in the beamformer design. We present several "intuitive" and optimal formulations of acoustic rake receivers both in the frequency domain and in the time domain, and show theoretically and numerically that the rake formulation of the maximum signal-to-interference-and-noise (Rake-MaxSINR) beamformer offers significant performance boosts in terms of noise and interference suppression. Beyond SINR, we observe gains in terms of the perceptual evaluation of speech quality (PESQ) metric for the speech quality.

Summary of Contributions in Chapter 6

- We propose the concept of an acoustic rake receiver (ARR) for speech—an echo-aware microphone beamformer. This is inspired by the spatial component of a well-known receiver design in wireless communications.
- We show theoretically and numerically that the Rake-Max-SINR formulation achieves large gains not only in terms of noise cancellation and interferer suppression as measured by the SINR, but also in terms of the PESQ metric, which means that our formulation

produces results that are perceptually superior to standard beamformers. This is corroborated by informal listening tests.

- We show that ARRs can separate the desired source from the interferer in situations where conventional beamformers are bound to fail; for example, if a strong interferer is occluding the direct path of the desired signal, an ARR will ignore the direct path and listen only to echoes.

Chapter 7: Alternative Generalized Inverses Similarly to Chapter 5, the final chapter of this thesis departs slightly from the central theme. It grew out of an effort to speed up certain overdetermined tomographic inversions related to a new touch display technology. The first approach was to use the Moore-Penrose pseudoinverse (MPP) of the system matrix in order to compute the reconstruction. But the demand on the frame rate was so stringent that applying the MPP was far too expensive in terms of computation time. The solution was to use an alternative generalized inverse with many zeros—the *sparse pseudoinverse*—obtained as the generalized inverse with the smallest entrywise ℓ^1-norm. Intrigued by the implications of this finding, we studied the concept of a norm-minimizing alternative generalized inverse in much more depth, and Chapter 7 describes this study.

A particular norm-minimizing generalized inverse is the MPP. It has the smallest Frobenius norm among all generalized inverses of a matrix. While the MPP is optimal well beyond the Frobenius norm, freeing up the degrees-of-freedom associated with the square-norm optimality enables us to enforce other useful properties. We first generalize the results of Zię̨tak [224] by showing that the MPP minimizes several norms beyond unitarily invariant norms, thus further establishing its *robustness* as the correct choice in most situations.

We then concentrate on some norms which are not minimized by the MPP, but whose minimization is relevant for linear inverse problems and sparse representations. In particular, we look at the entrywise ℓ^1-norm and the induced $\ell^p \to \ell^q$ norms. For example, we show how to compute generalized inverses that achieve a *poor man's ℓ^p minimization*, in the sense that they minimize the average-case or the worst-case ℓ^p-norm blowup. Furthermore, instead of norms, we concentrate on matrices with interesting behaviors. We exhibit a class of matrices for which the MPP minimizes norms that it generically does not minimize, and a class for which many norm-minimizing generalized inverses coincide, but not with the MPP. Finally, we discuss efficient computation of the generalized inverses associated to various norms.

Summary of Contributions in Chapter 6

- Motivated by the *sparse pseudoinverse* as an attractive alternative to the MPP from the point of view of computational efficiency, we propose a general notion of a *norm-minimizing* generalized inverse.
- We give a categorization of matrix norms with respect to whether they are minimized by the MPP, or they are generically not minimized by the MPP (further establishing the optimality of the MPP in various cases). We then exhibit several classes of matrices that behave irregularly; *e.g.*, matrices whose MPP minimizes norms that it generically does not minimize, or matrices for which many generalized inverses coincide, but not with the MPP.
- We show that all norm-minimizing generalized inverses are *unbiased* in the sense that the corresponding projection operator is on average a scaled identity over common random matrix ensembles. We also show that their Frobenius norm is well-behaved—an

important property for noise stability.

- We compute norm-minimizing generalized inverses that act as *poor man's* ℓ^p *minimizers* in the sense that they minimize the worst-case or the average-case ℓ^p-norm blowup.

Conclusion Within the conclusion, we embed a research proposal. Our toolbox for working with distances under noise, erasures, and unknown permutations is useful beyond room acoustics. Potential applications are in MIMO communications, autonomous robotic navigation and mapping, and depth imaging to mention a few. Perhaps the most exciting connection is that it could unlock various *ab initio* reconstruction methods in crystallography.

The wideband advantage could be used to design a new generation of *scattering microphones* that would specifically enhance source localization and blind source separation.

Chapter 2

On Echoes and Distances*

The term *echo* is often used to refer to reflections of sound waves. We use it generically to refer to reflections of any kind of wave, although in applications we discuss only sound and radio. To understand how echoes arise from propagating sound waves and how to model them, we start by a short account of the wave mechanics of sound. We then introduce the *image source model*: This model replaces walls (reflectors) by points, thereby simplifying the discussion from that of wave propagation in complex geometries, to that of points and distances.

This naturally brings us to the second key ingredient: Euclidean distance geometry. We exploit the geometry of the sets of points corresponding to acoustic sources, receivers and echoes by using the toolbox of the Euclidean distance matrices (EDMs). These simple tables of squared distances have surprisingly useful properties which we capitalize upon repeatedly.

There is a noticeable discontinuity between the first part on waves, and the second part on EDMs. In the first part we aim primarily at developing intuitions about the physics of waves and their reflections; this leads to a less dense discussion than that of the second part, in which (in addition to intuition) we develop concrete algorithms that we will use elsewhere in the thesis.

The material on waves is digested from classic texts on acoustics, room acoustics and PDEs: Morse and Ingard's *Theoretical Acoustics* [154], Kuttruff's *Room Acoustics* [117], Duffy's *Green's Functions with Applications* [67], Farlow's *Partial Differential Equations for Scientists and Engineers* [73] and Ihlenburg's *Finite Element Analysis of Acoustic Scattering* [94], and sprinkled with our observation and interpretations.

2.1 Undulating Particles: The Wave Mechanics of Sound and Radio

2.1.1 Mass on a Spring

It seems like a good idea to begin the study of echoes by studying simple oscillations. What is the connection between echoes and masses on springs? Echoes are reflections of waves, and waves consist of elementary oscillations. In the case of acoustic waves, particles of air oscillate around their equilibrium positions, causing the pressure to oscillate around its resting value—the atmospheric pressure.

---

*The material on distances in the second half of this chapter is part of our tutorial paper on EDMs [55]—a joint work with Reza Parhizkar, Juri Ranieri and Martin Vetterli.

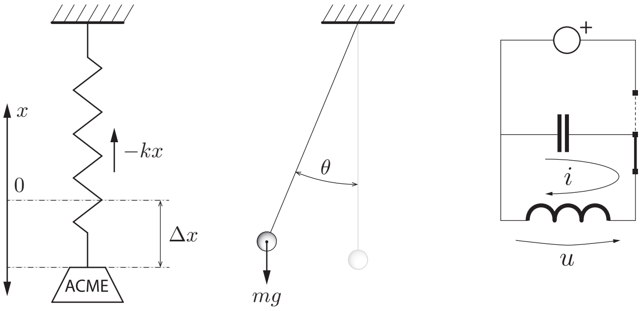


Figure 2.1: Examples of oscillating systems. The spring oscillates to counter stretching or compressing; the pendulum oscillates trying to return to equilibrium; the current (and the voltage) in a simple circuit oscillate.

Figure 2.1 shows three different oscillatory systems. Although they appear distinct, the equations that govern their behavior have the same form (linear second-order partial differential equation). We will go through the mechanics of the mass on a spring.

There are two prerequisites for oscillations: *inertia* and *stiffness*. Once the system is set in motion, inertia supports the motion past the equilibrium point and stiffness works to return the system back into the equilibrium. The two demands are balanced out by oscillatory motion, and a different kind of equilibrium: conservation of energy. In Figure 2.1, the stiffness component is provided by the spring that resists stretching, by the gravitational force that resists the pendulum moving up, and by the coil that resists the change of current (if we look at the oscillating current as opposed to the oscillating voltage). For the mass on the spring and the pendulum, inertia (first Newton's law) is provided by the mass that has maximal speed in what would otherwise be an equilibrium point, and by the capacitor which is fully charged when the current through the coil is zero, so it keeps pushing the current.

When a moving mass stretches the spring beyond its resting length, the kinetic energy of the mass is converted into the potential energy stored in the stretched spring. Once the mass runs out of kinetic energy, the motion is reversed and the spring gives the stored potential energy back to the mass. Without heat losses in the air or in the spring, the sum of the kinetic and the potential energy remains constant over time.

The spring is described by its stiffness, in our case a scalar k that tells us with how much force the spring will resist stretching by some unit length. The basic law of the spring—Hooke's law—is linear, so the force with which the spring resists stretching or compressing equals

$$F_s = -kx, \tag{2.1}$$

where x is the displacement from the equilibrium position, with reference to Figure 2.2. The negative sign indicates that the force acts in the direction opposite to the deformation.

We already mentioned that the law of the spring supplies the stiffness part of the stifness-inertia couple. The inertia part is played by the mass and Newton's first law, which says that an object in motion stays in motion with the same speed and direction as long as the total force on the object vanishes. When the spring is at its resting length, there are no forces acting on the mass, but it has a non-zero speed (in fact, the speed is maximal at this point), so it keeps moving. We can predict how the mass will move by invoking Newton's second law which says that the total force acting on a rigid body of mass m is proportional to the acceleration of the

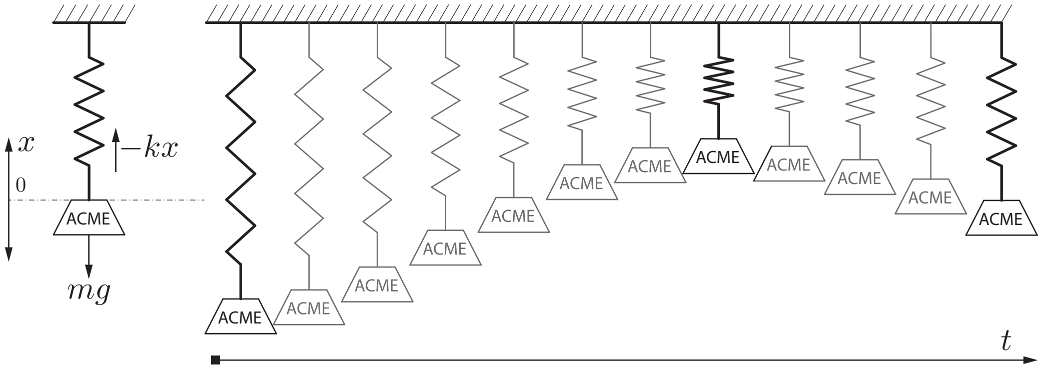


Figure 2.2: Oscillating mass on a spring.

body,

$$F = ma = m\frac{\mathrm{d}^2x}{\mathrm{d}t^2}. \tag{2.2}$$

In our case, the only force is the recoil force in the spring (for simplicity, we do not consider the gravitational force), so by combining (2.1) and (2.2), we get a differential equation that governs the motion of the mass—a second-order linear PDE with constant coefficients,

$$m\frac{\mathrm{d}^2x}{\mathrm{d}t^2} = -kx. \tag{2.3}$$

The solution to (2.3) is harmonic motion. It is easy to verify that the two linearly independent solutions are

$$x_1(t) = A\sin\left(\sqrt{\frac{k}{m}}t\right) \quad \text{and} \quad x_2(t) = B\cos\left(\sqrt{\frac{k}{m}}t\right). \tag{2.4}$$

Any solution can be expressed as a linear combination of these two, and any linear combination is a solution.

Example 2.1

Assume that at $t = 0$, the displacement of the mass is $x(0) = x_0$, and its velocity is $v(0) = v_0$. We want to compute the motion of the mass $x(t)$ for all t. The general solution to the mass-on-a-spring equation is

$$x(t) = A\sin(\omega_0 t) + B\cos(\omega_0 t),$$

for some value of the coefficients A and B, with $\omega_0 \overset{\text{def}}{=} \sqrt{k/m}$ being the frequency of oscillations in rad/s. We can find A and B using the initial conditions. We have that $x(0) = A\sin(0) + B\cos(0) = B = x_0$, and $x'(0) = A\omega_0\cos(0) - B\omega_0\sin(0) = A\omega_0 = v_0$. Finally we obtain

$$x(t) = \frac{v_0}{\omega_0}\sin(\omega_0 t) + x_0\cos(\omega_0 t) = \left[\left(\frac{v_0}{\omega_0}\right)^2 + x_0^2\right]^{1/2} \sin\left[\omega_0 t + \arctan\left(\frac{x_0\omega_0}{v_0}\right)\right],$$

for the motion of the mass.

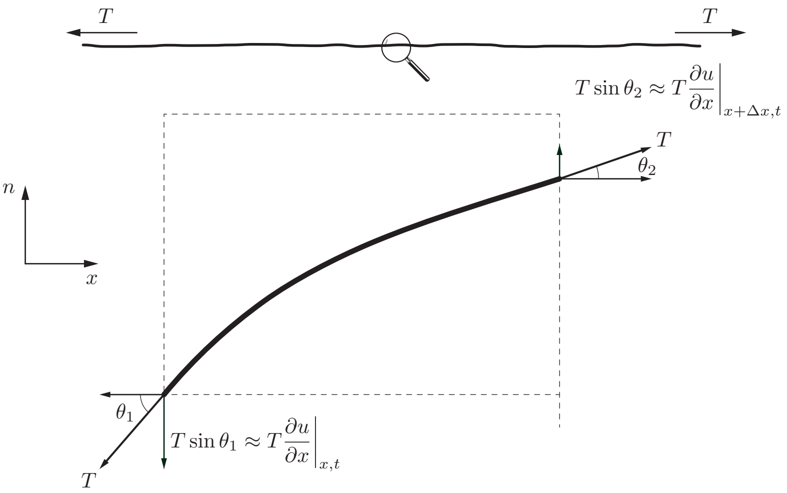

T
T
$T\sin\theta_2 \approx T\frac{\partial u}{\partial x}\Big|_{x+\Delta x,t}$
T
θ_2
n
x
θ_1
$T\sin\theta_1 \approx T\frac{\partial u}{\partial x}\Big|_{x,t}$
T

Figure 2.3: Forces on a segment of a vibrating string.

2.1.2 Vibrating String

The same laws and reasoning as for the harmonic oscillator are at play for a vibrating string. In fact, we could study the string as a system of coupled oscillators—masses connected with springs—but that would obfuscate an important analysis principle: we should be seeing the string as a whole, having a definite shape at any given time instant. As Morse and Ingard [154] put it, *"The motion of the string at any instant will depend on the shape of the string, and the subsequent shape will depend on the motion; what we must do is to find the relation between the shape and the motion"*. This relation for the vibrating string is a consequence of Newton's laws. Conveniently, the same principle applies for vibrations of a membrane and vibrations of air.

Consider an infinitely long string pulled taut by a tension force T. Let us concentrate on the segment between x and $x + \Delta x$ as illustrated in Figure 2.3, and write out the forces acting on the segment. The wave equation will simply be Newton's equation of motion applied to all such segments at once.

Waves on a string are *transversal.* That is, for oscillations of small amplitude, the velocity of a fixed segment of the string has no component parallel with the string. We are therefore interested in forces that act in the direction perpendicular to the string. With reference to Figure 2.3, the principal force that governs the string's behavior is the tension force. The net tension force in the transverse direction is

$$T_v = T\sin\theta_2 - T\sin\theta_1 \approx T\left(\frac{\partial u}{\partial t}\Big|_{x+\Delta x,t} - \frac{\partial u}{\partial t}\Big|_{x,t}\right), \tag{2.5}$$

where the approximation is valid as long as θ_1 and θ_2 are small (this is true when the oscillations are small). We can anticipate that the tension force will contribute the $\frac{\partial^2 u}{\partial x^2}$ term to the putative wave equation. Other forces include the external driving force $F(x,t)$ (for example the gravity mg, or impulses on the string). Another important force is friction, for example against the medium (air), equal to $-\beta\frac{\partial u}{\partial t}$. We could add more forces making the model more and more realistic, but this is not necessary to understand the principles.

Assume that the linear density of the string is ϱ. Then by applying Newton's equation of motion (second law, $F = ma$) to the segment, we get

$$\varrho \Delta x \frac{\partial^2 u(x,t)}{\partial t^2} = T \left[\left.\frac{\partial u}{\partial x}\right|_{(x+\Delta x,t)} - \left.\frac{\partial u}{\partial x}\right|_{(x,t)} \right] + \Delta x F(x,t) - \Delta x \beta \frac{\partial u(x,t)}{\partial t}. \tag{2.6}$$

Dividing both sides by Δx, and taking the limit as Δx goes to zero, we get the wave equation with damping (the telephone equation).

$$\frac{\partial^2 u}{\partial t^2} = \alpha^2 \frac{\partial^2 u}{\partial x^2} - \frac{\beta}{\varrho}\frac{\partial u}{\partial t} + \frac{1}{\varrho}F(x,t). \tag{2.7}$$

Let us now take a closer look at the simple wave equation without friction and sources,

$$\frac{\partial^2 u}{\partial t^2} = \alpha^2 \frac{\partial^2 u}{\partial x^2}. \tag{2.8}$$

Intuitive Solution We know from experience that waves on strings and ropes can travel. If we rapidly pull one end of a rope up and down, the disturbance will travel along it until it reaches the other end (where it may reflect, but more on that later). The shape of this disturbance remains almost constant as it travels. Imagine that the shape of the string at $t = 0$ is $u_1(x)$, and that it represents a wave traveling to the right. After a time t, it will have moved to the right by a distance ct, where c is the speed of the wave, so the new shape will be $u_1(x - ct)$. Similarly, for a perfect motion of the wave to the left, the shape of the wave should be $u_2(x + ct)$ at time t, given that it is $u_2(x)$ at time zero. The equation of motion is linear, so the general expression for a perfect wave motion is the sum (actually a linear combination) of the two waves,

$$u(x,t) = u_1(x - ct) + u_2(x + ct). \tag{2.9}$$

To check if (2.9) is really a solution, we can take partial derivatives of u. The partial derivative $\partial u/\partial x$ represents the slope of the curved string at a given time and position; the time derivative $\partial u/\partial t$ is the velocity in the transverse direction. For a right traveling wave, we see from (2.9) that the two partial derivatives are related as

$$\frac{\partial u}{\partial t} = -c \left.\frac{\partial u_1}{\partial t}\right|_{x-ct} = -c\frac{\partial u}{\partial x}. \tag{2.10}$$

Similarly, for the left traveling wave the relation is $\frac{\partial u}{\partial t} = c\frac{\partial u}{\partial x}$. We can differentiate (2.10) to see that both the left and the right traveling wave satisfy the second-order equation

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}. \tag{2.11}$$

But this is exactly the wave equation (2.8) with $\alpha = c$. Thus we have identified the physical meaning of the coefficient α.

The above derivation shows that (2.9) is definitely *a* solution to (2.11). In fact, it can be shown that the solution of the form (2.9) implies the wave equation (2.11) and vice versa, so (2.11) can indeed be called the wave equation, and any solution has the form (2.9).

Canonical Coordinates A more formal approach to solve (2.11) is to introduce (now aided by hindsight) the *canonical coordinates* (ξ, η) as

$$\begin{aligned}\xi &= x + ct\\ \eta &= x - ct.\end{aligned}$$

It is not difficult to see that (2.11) in the canonical coordinates becomes

$$\frac{\partial^2 u}{\partial \xi\, \partial \eta} = 0. \tag{2.12}$$

Integration with respect to ξ and then with respect to η gives

$$u(\xi, \eta) = \phi(\eta) + \psi(\xi), \tag{2.13}$$

where $\phi(\eta)$ is an arbitrary function of η, and $\psi(\xi)$ an arbitrary function of ξ. Transforming back to the original coordinates, we get

$$u(x, t) = \phi(x - ct) + \psi(x + ct). \tag{2.14}$$

Both derivations show that in any solution of the wave equation on a string, time and space appear laced up together as $x - ct$ or $x + ct$. It may not be explicit in a particular expression, but there *must* be an equivalent expression where it is. It is interesting that a sum of *any* two traveling waves—one traveling right, the other one traveling left—is a solution of the wave equation on a string.

Example 2.2

All of the following functions are solutions of the wave equation on a string (2.11) *(by* tri(x) *we denote the unit triangle pulse):*

(a) $u_1(x, t) = \mathrm{tri}(x - ct)$,

(b) $u_2(x, t) = \mathrm{rect}(x - ct) - \mathrm{tri}(x + ct)$,

(c) $u_3(x, t) = \sin(x)\cos(ct) = \frac{1}{2}\sin(x - ct) + \frac{1}{2}\sin(x + ct)$,

(d) $u_4(x, t) = \mathrm{tri}(2(x - ct) + 6.6) - \frac{1}{2}\big[1 + \cos(2\pi(x + ct - 3.3)\big]\mathrm{rect}(x + ct - 3.3)$.

Figure 2.4 shows the development of $u_4(x, t)$ *over time.*

2.1.3 Wave Equation of Sound

Oscillations of air particles around their equilibrium positions create small pressure changes around the atmospheric pressure which are then propagated as a wave. We call this wave *sound*, in particular when we can hear it.

As we will see, a major difference between the sound waves and the waves on a string or on a membrane is that the sound waves are *longitudinal*, not transversal. This means that the particles of air oscillate along the direction of wave propagation, not transversally to it, as illustrated in Figure 2.5. A membrane oscillates up and down about the equilibrium plane; in air we have compressions and rarefactions.

We start by writing out the linearized equations for various quantities associated with compressible fluids, and then combine them using Newton's law of motion into the acoustic wave equation.

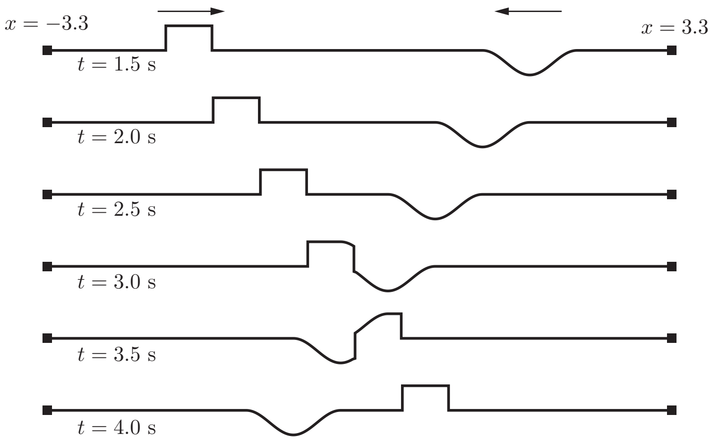


Figure 2.4: Example of traveling waves on a string.

Conservation of Mass Consider a fluid whose state is described by pressure $P(\boldsymbol{x}, t)$, density $\rho(\boldsymbol{x}, t)$, and particle velocity $\boldsymbol{v}(\boldsymbol{x}, t)$. Conservation of mass means that the only way the mass of the fluid inside a volume V can change is through a net inflow or outflow of the fluid into V through the boundary ∂V (Figure 2.6),

$$-\frac{\partial}{\partial t}\int_V \varrho \, \mathrm{d}V = \oint_{\partial V} \varrho \langle \boldsymbol{v}, \boldsymbol{n} \rangle \, \mathrm{d}S. \tag{2.15}$$

The left hand side of (2.15) equals the temporal change of the total mass inside V; the right hand side equals the net flow of the mass throught the boundary ∂V. The two terms must match. Using the Green-Gauss-Ostrogradski divergence theorem , we transform the integral on the right-hand side into a volume integral,

$$\oint_{\partial V} \varrho \langle \boldsymbol{v}, \boldsymbol{n} \rangle \, \mathrm{d}S = \int_V \nabla(\varrho \boldsymbol{v}) \, \mathrm{d}V, \tag{2.16}$$

yielding finally

$$\int_V \left(\frac{\partial \varrho}{\partial t} + \nabla(\varrho \boldsymbol{v}) \right) \, \mathrm{d}V = 0. \tag{2.17}$$

As this must hold no matter how we choose the volume V, it follows that

$$\frac{\partial \varrho}{\partial t} + \nabla(\varrho \boldsymbol{v}) = 0. \tag{2.18}$$

Equation of Motion The net force on the volume of air V resulting from the pressure $p(\boldsymbol{x}, t)$ is

$$\boldsymbol{F} = -\oint_{\partial V} p\boldsymbol{n} \, \mathrm{d}S.$$

Same as in the case of an oscillator or a vibrating string, we now use Newton's second law, obtaining

$$-\oint_{\partial V} p\boldsymbol{n} \, \mathrm{d}S = \int_V \varrho \frac{\mathrm{d}\boldsymbol{v}}{\mathrm{d}t} \, \mathrm{d}V. \tag{2.19}$$

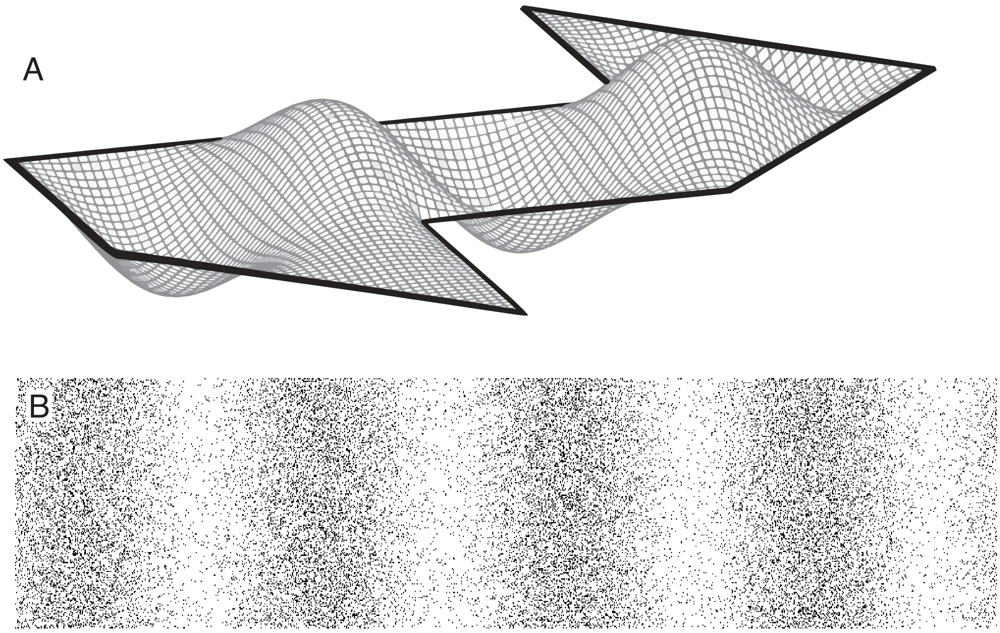


Figure 2.5: Illustration of transversal (A) and longitudinal (B) waves.

The total differential $\mathrm{d}\boldsymbol{v}/\mathrm{d}t = \partial\boldsymbol{v}/\partial t + \langle \boldsymbol{v}, \nabla \rangle\, \boldsymbol{v}$ is linearized by neglecting the second-order term, because in acoustics, and especially in room acoustics, we assume that the oscillations are small (*i.e.*, the elementary oscillators operate in the linear regime). With another application of the Green-Gauss-Ostrogradski theorem we get the equation of motion, known as Euler's equation or the law of conservation of momentum [117, 154],

$$\varrho \frac{\partial \boldsymbol{v}}{\partial t} = -\nabla p. \tag{2.20}$$

The equation of conservation of momentum (2.20) reveals an important fact: The air particles oscillate only along the direction of the pressure gradient, which is incidentally also the direction of wave propagation. This confirms what we mentioned earlier—the sound waves are longitudinal, in contrast to waves on strings and membranes.

The particle velocity $\boldsymbol{v}$ is the time derivative of the particle position (displacement) $\boldsymbol{u}$, so we can rewrite Euler's equation as

$$\varrho \frac{\partial^2 \boldsymbol{u}}{\partial t^2} = -\nabla p. \tag{2.21}$$

Combining into Wave Equation Similarly as with the simple oscillator, we assume a linear material law,

$$p = c^2 \varrho. \tag{2.22}$$

In words, the denser the air, the higher the pressure. The square root of the proportionality constant, c, is the speed of sound. Instead of simply anticipating it, we could have deduced it from the final wave equation similarly to how we did it for the string.

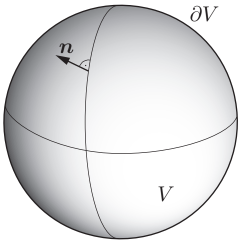


Figure 2.6: Volume element for reference in the wave equation derivation.

Combining with (2.18) and (2.20), we get

$$\frac{\partial^2 p}{\partial t^2} = c^2 \frac{\partial^2 \varrho}{\partial t^2} = -c^2 \varrho_0 \nabla \frac{\partial \boldsymbol{v}}{\partial t} = c^2 \nabla(\nabla p), \tag{2.23}$$

where we first linearized the density ρ around the equilibrium value ρ_0 and neglected the higher order terms. Rearranging we finally obtain the wave equation,

$$\Delta p - \frac{1}{c^2} \frac{\partial^2 p}{\partial t^2} = 0. \tag{2.24}$$

The wave equation of sound (2.24) is completely analogous to the wave equation on a string (2.11). The only difference is that the second partial derivative with respect to the spatial variable is replaced by the sum of the partial derivatives with respect to all three (or two) spatial variables—the Laplacian Δ.

If we assume that the waves are time-harmonic, that is, $p(\boldsymbol{x}, t) = \widehat{p}(\boldsymbol{x}) \mathrm{e}^{\mathrm{j}\omega t}$, we get the Helmholtz equation,

$$\Delta \widehat{p} + k^2 \, \widehat{p} = 0, \tag{2.25}$$

where $k \overset{\text{def}}{=} \omega / c$ is the *wavenumber* (the number of waves in 1 m). Note that the Helmholtz equation can also be obtained by taking the continuous-time Fourier transform of the wave equation (2.24).

2.1.4 Plane-wave Solution

We can obtain a solution to the wave equation of sound similar to the solution for waves on a string (2.14) by assuming that the pressure changes only along one direction (the direction of the velocity vector $\boldsymbol{c}$) not only locally, but in the whole space. This results in plane waves, defined as the waves whose wavefronts—surfaces of constant phase—are parallel planes, normal to the direction of the wave propagation. Note that on a string, there are only two directions in which any wave can travel: x and $-x$; in air, there are infinitely many directions.

We can guess the general expression for a plane wave $p(\boldsymbol{x}, t)$ with the velocity $\boldsymbol{c}$ by considering the following two requirements:

(i) For a fixed time t_0 the pressure at all the points in any plane perpendicular to $\boldsymbol{c}$ must be equal;

(ii) At time $t_0 + t$, the plane containing the point $\boldsymbol{x}$ must have moved to $\boldsymbol{x} + t\boldsymbol{c}$.

The first requirement is satisfied by functions of the form $\phi(\langle \boldsymbol{x}, \widehat{\boldsymbol{c}} \rangle)$, where by $\widehat{\boldsymbol{c}}$ we denote the unit vector corresponding to $\boldsymbol{c}$, $\boldsymbol{c}/\|\boldsymbol{c}\|$. The second requirement suggest that we must find α so that

$$\langle \widehat{\boldsymbol{c}}, \boldsymbol{x} + \boldsymbol{c} \rangle + \alpha \tau = \langle \widehat{\boldsymbol{c}}, \boldsymbol{x} \rangle,$$

where $c = \|\boldsymbol{c}\|$. It follows that $\alpha = -c$, so that the plane wave solution has the form

$$p(\boldsymbol{x}, t) = \phi(\langle \boldsymbol{x}, \widehat{\boldsymbol{c}} \rangle - ct). \tag{2.26}$$

It is straightforward to verify that this function satisfies the acoustic wave equation.

Planar wavefronts can be seen as an approximation of spherical wavefronts far away from the point source that generated the spherical wave. Indeed, in situations where point sources are *far enough*, the exact spherical solution is commonly replaced by plane waves; we say that the sources are in the *far field*.

It is much more common to define plane waves for the time-harmonic case. If we choose $\phi(\xi) = \mathrm{e}^{\mathrm{j}\omega\xi/c}$, we get

$$p(\boldsymbol{x}, t) = \mathrm{e}^{\mathrm{j}(\langle \boldsymbol{k}, \boldsymbol{x} \rangle - \omega c)}. \tag{2.27}$$

That is, the time-harmonic plane waves are also space-harmonic.

2.2 Origin of Echoes

2.2.1 Boundary Conditions in Wave Equations

For the wave equation to have a unique solution, we need to impose certain boundary conditions. In rooms, the boundary conditions describe the reflective properties of the walls; in free space, the boundary conditions are a limiting process—they describe the wave's behavior far away from the sources.

Here, we are only interested in boundary conditions that correspond to actual boundaries, typically to walls. There are two natural extremes and the corresponding logic is different for waves on a string and for waves in air. If one end of the string is attached to a rigid support, then the boundary condition is that this end of the string cannot move, so that the *displacement* at this end is zero.

In air, we usually write the wave equation for the pressure, not for the displacement. But it is clear that the pressure variation at the wall cannot be zero. We will see that a quantity that *does* vanish at the wall is the pressure gradient in the normal direction.

The first kind of boundary condition where the function itself is zero at the boundary is called Dirichlet (or *soft* wall, pressure does not vary at the wall). The second kind of boundary condition where the derivative is zero is called Neumann (or *hard* wall, pressure gradient vanishes at the wall).

2.2.2 Reflections of Ropes

Let us take a closer look on how reflections are formed on a string. For simplicity, we assume that the end of the string corresponding to $x = 0$ is fastened down for a perfectly rigid support. The boundary condition is then simply that this end cannot move,

$$u(0, t) = 0. \tag{2.28}$$

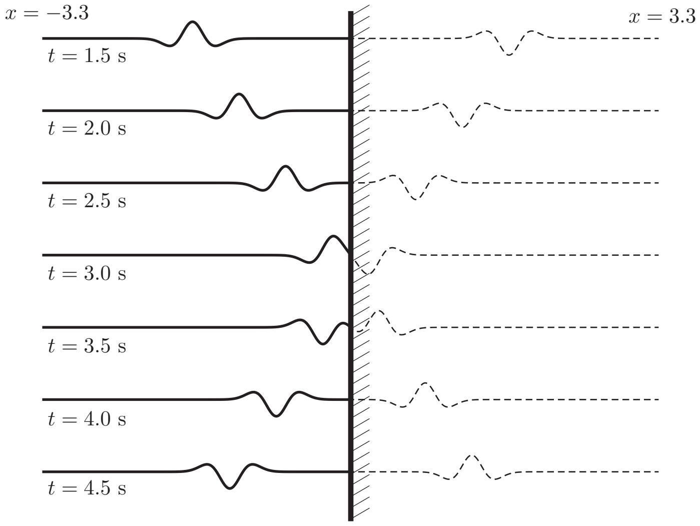


Figure 2.7: Reflections on a string. The non-phyiscal part of the string that carries the inverted wave is shown dashed.

The idea is to find the correct left- and right-traveling waves so that (2.28) is always true. It is easy to see that taking only $u_1(x-ct)$ or $u_2(x+ct)$ will not make the cut, as these waves will at some point pass the point $x = 0$, and they will displace the string there. But the following simple combination works:

$$u(x,t) = -f(x - ct) + f(-x - ct), \tag{2.29}$$

for at $x = 0$, and for all t we have

$$u(0,t) = -f(-ct) + f(-ct) = 0. \tag{2.30}$$

The form of the solution (2.29) has for a consequence the familiar behavior of the phase of waves on a string or on a rope; that is, the reflected wave is inverted.

A valid and important question is whether we have exploited all solutions this way. The answer is yes: A Dirichlet problem that we just described is known to have a unique solution. Since we have already found one, we need not search further.

Example 2.3

An example of a wave on a string reflected from a rigid support is given in Figure 2.7. The waveform of the pulse is approximately equal to

$$\phi(x) = \frac{\mathrm{d}^2}{\mathrm{d}x^2} \exp(-25x^2)$$

when the pulse is far enough from the support (due to the fast decay of the function). The exact waveform on the string (for a support located at $x = 0$*) is*

$$u(x,t) = \phi(x - c(t-\tau)) - \phi(-x - c(t-\tau)),$$

where τ *is the time offset.*

Typically, we would start from the initial conditions, $u(x,0)$ *and* $\frac{\partial u}{\partial t}\big|_{t=0}$*, and work to compute* u *for every* x *and* t*. One way to do it for traveling waves would be to identify the wave velocity* c*.*

2.2.3 Reflections of Sound

Let us go through a phenomenological explanation of sound wave reflections [117, 154], and then verify that the solution obtained by this observation is also compatible with the requested boundary conditions.

A plane wave incident at a flat wall of infinite extent reflects from the wall. The wall is characterized by its reflection factor R that determines how the amplitude and the phase of the reflected wave are altered relative to the incident wave,

$$R = |R|\mathrm{e}^{\mathrm{j}\chi}. \tag{2.31}$$

The intensity of the incident wave is proportional to the square of its amplitude, so we can also define the wall's *absorption factor* α,

$$\alpha = 1 - |R|^2. \tag{2.32}$$

Both $R = +1$ and $R = -1$ yield $\alpha = 0$, that is, no energy is absorbed by the wall. But these two extreme cases define physically very different scenarios. When $R = +1$, the reflected wave is in phase with the incident wave, and the wall is said to be sound-hard. By hard, it is meant that it does not yield to the oscillating pressure at it. On the other hand, $R = -1$ characterizes a sound-soft wall. Loosely speaking, soft walls move *in sync* with the oscillating particles, so that they feel no pressure.

A quantity related to the reflection factor is the wall impedance. It is defined as the ratio of the pressure and the normal component of the particle velocity at the wall,

$$Z = \left(\frac{p}{v_n}\right)_{\text{@surface}}. \tag{2.33}$$

The wall impedance plays the same role for the pressure and the particle velocity as the electric impedance plays for the voltage and the current. It is usually defined for sound waves of a particular frequency, when they are expressed as phasors (*i.e.*, not as sines and cosines, but as complex exponentials). It is therefore a complex quantity.

Normal Incidence For simplicity, we start with a plane wave that arrives at the wall at a normal incidence. The wall is assumed to correspond to the $x = 0$ coordinate plane. We also make a simplifying assumption that the wave is time-harmonic,

$$p_i(x, t) = p_0 \, \mathrm{e}^{\mathrm{j}(\omega t - kx)}. \tag{2.34}$$

Then the velocity of the oscillating particles is also time-harmonic (*cf.* conservation of momentum law (2.20))

$$v_i = \frac{p_0}{\varrho_0 c} \mathrm{e}^{\mathrm{j}(\omega t - kx)}. \tag{2.35}$$

The amplitude and the phase of the reflected wave are modified according to the magnitude and the phase of R. Reasons of symmetry suggest that the reflected wave will travel in the direction opposite from the incident wave. The change in the direction of travel means that we must change the sign of k. We thus obtain

$$p_r(x, t) = R p_0 \mathrm{e}^{\mathrm{j}(\omega t + kx)} \tag{2.36}$$

$$v_r(x, t) = -R \frac{p_0}{\varrho_0 c} \mathrm{e}^{\mathrm{j}(\omega t + kx)}. \tag{2.37}$$

This reflected wave is superposed to the incident wave, and they interfere. The total sound pressure and particle velocity in the wall plane are

$$p(0, t) = p_0 (1 + R) \mathrm{e}^{\mathrm{j}\omega t} \tag{2.38}$$

$$v(0, t) = \frac{p_0}{\varrho c} (1 - R) \mathrm{e}^{\mathrm{j}\omega t}. \tag{2.39}$$

To compute the wall impedance, we divide the pressure by the normal component of the particle velocity at the wall. Since the wave is arriving perpendicularly at the wall and the sound waves are longitudinal, the particle velocity only has the normal component so that the impedance is

$$Z = \varrho_0 c \frac{1 + R}{1 - R}. \tag{2.40}$$

Note that there is a satisfying parallelism between the electric and the acoustic impedances. A rigid wall with $R = 1$ always has zero velocity regardless of the pressure of particles around it; its impedance is thus $Z = \infty$. Current through an open circuit is always 0, regardless of the voltage across it—its electric impedance is ∞. For a soft wall with $R = -1$ (short circuit), the impedance vanishes. For a wall that perfectly absorbs the acoustic energy, the impedance equals the impedance of the air (matching). Interestingly both $Z = 0$ and $Z = \infty$ generate reflections (of opposite phase), but a wall with $Z = \varrho_0 c$ does not generate a reflection. This situation is analogous to *impedance matching* on transmission lines, where a load of matching impedance does not generate a reflection, but both open and short circuit do.

Oblique Incidence If the wave is arriving at an angle θ with respect to the wall normal, as in Figure 2.8, it is traveling in the direction $\boldsymbol{n} = [\cos\theta, \ \sin\theta]^\top$. The plane wave phasor (without the time-dependent part) is given as

$$p_i = p_0 \mathrm{e}^{-\mathrm{j}k\langle \boldsymbol{x}, \boldsymbol{n} \rangle} = p_0 \mathrm{e}^{-\mathrm{j}k(x \cos\theta + y \sin\theta)}. \tag{2.41}$$

Let us compute the wall impedance. The normal component of the particle velocity is given for the time-harmonic wave as

$$v_x = -\frac{1}{\mathrm{j} w \varrho_0} \frac{\partial p}{\partial x}, \tag{2.42}$$

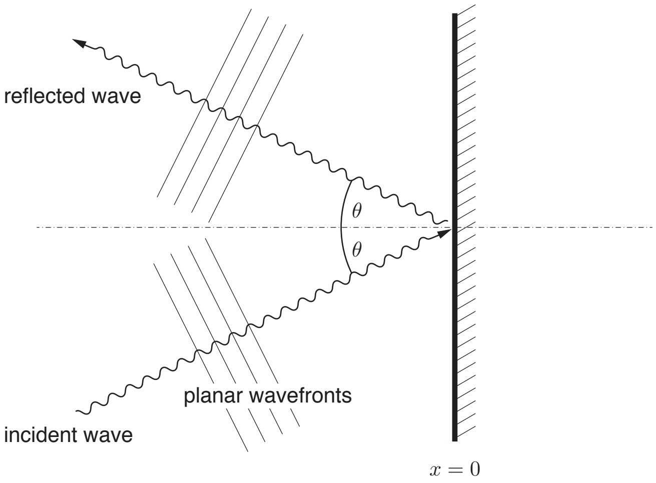


Figure 2.8: Illustration of the reflected wave at an oblique incidence.

so for the incident wave (2.41), we get

$$(v_i)_x = \frac{p_0}{\varrho_0 c}\cos(\theta)\mathrm{e}^{-\mathrm{j}k(x\cos\theta + y\sin\theta)}. \tag{2.43}$$

Due to the way we set up the coordinate system, for the reflected wave we must reverse the sign of the x component of $\boldsymbol{n}$, or equivalently, of the wave vector $\boldsymbol{k}$. Pressure and particle velocity of the reflected wave are then

$$p_r = Rp_0\mathrm{e}^{-\mathrm{j}k(-x\cos\theta + y\sin\theta)} \tag{2.44}$$

$$(v_r)_x = -R\frac{p_0}{\varrho_0 c}\cos(\theta)\mathrm{e}^{-\mathrm{j}k(-x\cos\theta + y\sin\theta)}. \tag{2.45}$$

After computing the total pressure and particle velocity (incident + reflected), we get the wall impedance

$$Z = \frac{\varrho_0 c}{\cos\theta}\frac{1+R}{1-R}. \tag{2.46}$$

We can see that the wall impedance changes with the angle of incidence θ.

2.3 Room Acoustics and Echo Modeling

2.3.1 Green's Function

Green's function is one of the central tools in analysis of linear ODEs and PDEs. It is a natural generalization of the *impulse response* from signals and systems. Consider a linear differential equation

$$\begin{cases} Df(\boldsymbol{x}, t) = s(\boldsymbol{x}, t) \\ \text{Boundary conditions,} \end{cases} \tag{2.47}$$

where $s(\boldsymbol{x}, t)$ is the source term. Then the Green's function is the solution of (2.47) when the source is a perfect instantaneous point source,

$$s(\boldsymbol{x}, t) = \delta(\boldsymbol{x} - \boldsymbol{x}_s)\delta(t - t_s),$$

where δ is an adequately defined Dirac delta distribution.

The solution for any source term with arbitrary spatial and temporal dependency is then computed as a convolution of the source term with the Green's function,

$$f(\boldsymbol{x}, t) = \iint g(\boldsymbol{x}, t \,|\, \boldsymbol{\xi}, \tau) s(\boldsymbol{\xi}, \tau) \mathrm{d}\boldsymbol{\xi} \ \mathrm{d}\tau. \tag{2.48}$$

The Green's function is a complete description of a linear physical system, in the same way the impulse response is a complete description of a linear system. Impulse response is typically defined for time-invariant linear systems, but there is no problem with defining for time-variant systems as well. Green's functions that we will work with are time-invariant too, but they are spatially varying. Therefore, the convolution over time in (2.48) will be the "ordinary" time-invariant convolution

Green's Function for the Wave Equation Because the final result is so well-known and commonly used, we now go through the details of the computation of the Green's function for the wave equation in free space. We follow the derivation of Duffy [67]. The problem to solve is

$$\Delta g - \frac{1}{c^2}\frac{\partial^2 g}{\partial t^2} = \delta(\boldsymbol{x} - \boldsymbol{\xi})\delta(t - \tau), \tag{2.49}$$

with the initial condition

$$g(\boldsymbol{x}, 0 \,|\, \boldsymbol{\xi}, \tau) = g_t(\boldsymbol{x}, 0 \,|\, \boldsymbol{\xi}, \tau). \tag{2.50}$$

We first take the Laplace transform of (2.49) (with respect to time),

$$\Delta G - \frac{s^2}{c^2} G = -\delta(\boldsymbol{x} - \boldsymbol{\xi}) \mathrm{e}^{-s\tau}, \tag{2.51}$$

and then introduce the inverse 3D CFT over space/wave vector,

$$G(\boldsymbol{x}, s \,|\, \boldsymbol{\xi}, \tau) = \frac{1}{(2\pi)^3} \iiint\limits_{\mathbb{R}^3} \widehat{G}(\boldsymbol{k}, s \,|\, \boldsymbol{\xi}, \tau) \mathrm{e}^{\mathrm{j}\langle \boldsymbol{k}, \boldsymbol{x} \rangle} \ \mathrm{d}\boldsymbol{k}. \tag{2.52}$$

Substituting (2.52) into (2.51), we get

$$\widehat{G}(\boldsymbol{k}, s \,|\, \boldsymbol{\xi}, \tau) = \frac{c^2}{s^2 + c^2 \kappa^2} \mathrm{e}^{-\mathrm{j}\langle \boldsymbol{k}, \boldsymbol{\xi} \rangle - s\tau}. \tag{2.53}$$

Finally,

$$G(\boldsymbol{x}, s \,|\, \boldsymbol{\xi}, \tau) = \frac{\mathrm{e}^{-s\tau}}{(2\pi)^3} \iiint\limits_{\mathbb{R}^3} \frac{\mathrm{e}^{\mathrm{j}\langle \boldsymbol{k}, \boldsymbol{x} - \boldsymbol{\xi} \rangle}}{\kappa^2 + s^2/c^2} \ \mathrm{d}\boldsymbol{k}. \tag{2.54}$$

To solve (2.54), we set up a spherical coordinate system (r, θ, ϕ) so that $\boldsymbol{x} - \boldsymbol{\xi} = [0,\ 0,\ r]^\top$. We obtain that

$$G(\boldsymbol{x}, s \,|\, \boldsymbol{\xi}, \tau) = \frac{\mathrm{e}^{-s(r/c - \tau)}}{4\pi r}. \tag{2.55}$$

Taking the inverse Laplace transform we get

$$g(\boldsymbol{x}, t \,|\, \boldsymbol{\xi}, \tau) = \frac{\delta(t - \tau - r/c)}{4\pi r}. \tag{2.56}$$

This is the famous solution for the point source. You can imagine that sound exists only within an infinitesimally thin spherical shell whose radius is growing at the speed of sound. A listener at a distance r from the sound source will hear the pulse after r/c, and then never again. It is interesting to contrast this with the solution for the 2D wave equation in the plane that can be obtained by a similar procedure,

$$g_{\mathrm{2D}}(\boldsymbol{x}, t \,|\, \boldsymbol{\xi}, \tau) = \frac{1}{2\pi} \frac{\mu(t - \tau - r/c)}{\sqrt{(t-\tau)^2 - (r/c)^2}}, \tag{2.57}$$

where μ is the Heaviside step function. In the flatland, the sound intensity decays, but the sound never really goes away.

2.3.2 Image Source Method

The image sources (IS) method is a very well-known tool in acoustics. It belongs in the category of geometrical acoustics, and in many texts it is introduced phenomenologically. Allen and Berkley wrote a paper in 1979 [5] that became *the* reference on the IS method, and it was extended to polyhedral geometries beyond shoe-box by Borish [19]. But the method itself was known well before; it is a standard technique to solve PDEs on certain domains. The best way to see how that works is by using Green's functions.

The free-space Green's function is a *particular* solution of any Green's function problem. For this reason, it is sometimes called the *fundamental solution*. When we introduce boundary conditions[1] into the problem, the fundamental solution still satisfies the PDE everywhere except at the boundaries, as it does not "know" about the boundary conditions. That is why when solving a PDE we also search for a *homogeneous* solution such that the sum of the particular and the homogeneous solution *does* satisfy the boundary conditions.

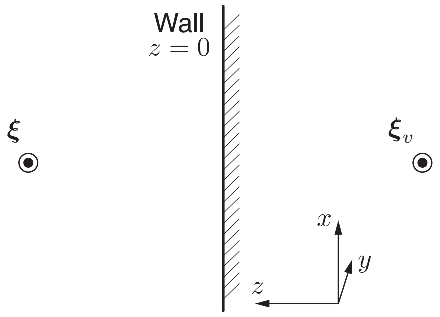


Figure 2.9: Construction of the Green's function for the Helmholtz equation in a halfspace.

---

[1] Other than the so-called radiation condition which specifies the wave's behavior at infinity.

Green's Function in a Halfspace Let us find the Green's function for the Helmholtz equation in a 3D halfspace $z > 0$, as illustrated in Figure 2.9. We will need the fundamental solution for the Helmholtz equation, given as [67]

$$g_0(\boldsymbol{x} \,|\, \boldsymbol{\xi}) = \frac{\mathrm{e}^{\mathrm{j}k\|\boldsymbol{x}-\boldsymbol{\xi}\|}}{4\pi\|\boldsymbol{x}-\boldsymbol{\xi}\|}. \tag{2.58}$$

Note that we can obtain this solution by taking the Fourier transform of the Green's function for the wave equation (2.56).

We assume that the infinite wall $\mathcal{Z}$ at $z = 0$ is perfectly reflective. The problem that we have to solve is then

$$\begin{gathered} -\Delta g - k^2 g = \delta(\boldsymbol{x}-\boldsymbol{\xi}) \\ \langle \nabla g(\boldsymbol{x} \,|\, \boldsymbol{\xi}), \boldsymbol{n}_z \rangle = 0 \text{ for } \boldsymbol{x} \in \mathcal{Z}, \end{gathered} \tag{2.59}$$

where $\boldsymbol{n}_z = [0,0,1]^\top$ is the boundary (wall) normal, equivalent to the hard wall boundary in the wave equation. For the free-space Green's function g_0, the value of the gradient's normal component at the wall is not zero; we have to make it so. We do it in two steps.

First, we recall that because of the linearity of the Helmholtz equation and the properties of the fundamental solution, the equation (2.59) is satisfied by any linear combination of fundamental solutions $\sum_{i=1}^n g_0(\boldsymbol{x} \,|\, \boldsymbol{\xi}_i)$ everywhere except at the boundary, and at the points $\boldsymbol{\xi}_i$.

The second step is to place another, *virtual* source so that the boundary condition be satisfied. This turns out to be an easy task: Symmetry suggests that a virtual source at the symmetric position of $\boldsymbol{\xi}$ will create the same pressure on the wall from the opposite side, so that the gradient of the pressure will be zero (the pressure itself will double, and we know that this is the correct behavior). We can verify this analytically,

$$g(\boldsymbol{x} \,|\, \boldsymbol{\xi}) = g_0(\boldsymbol{x} \,|\, \boldsymbol{\xi}) + g_0(\boldsymbol{x} \,|\, \boldsymbol{\xi}_v) = \frac{\mathrm{e}^{\mathrm{j}k_0\sqrt{(z-\zeta)^2+\varrho^2}}}{4\pi\sqrt{(z-\zeta)^2+\varrho^2}} + \frac{\mathrm{e}^{\mathrm{j}k_0\sqrt{(z+\zeta)^2+\varrho^2}}}{4\pi\sqrt{(z+\zeta)^2+\varrho^2}}, \tag{2.60}$$

where $\boldsymbol{x} = [x,y,z]^\top$, $\boldsymbol{\xi} = [\xi,\eta,\zeta]^\top$, $\varrho^2 = (x-\xi)^2 + (y-\eta)^2$, and

$$\boldsymbol{\xi}_v = \boldsymbol{\xi} + 2\langle \boldsymbol{\xi}, \boldsymbol{n}_z \rangle \boldsymbol{n}_z$$

is the mirror image of $\boldsymbol{\xi}$ in the plane $z = 0$. The first term in (2.60) is the free-space Green's function g_0, so it satisfies the Helmholtz equation with the source term $\delta(\boldsymbol{x}-\boldsymbol{\xi})$. The second term is the fundamental solution for a virtual mirror source. Plugging in $z = 0$, we see that we get twice the pressure that would be there without the wall. Furthermore, it satisfies the homogeneous equation.

To verify that the normal pressure gradient at the wall is zero, we only need to look at the partial derivative with respect to the z coordinate:

$$\frac{\partial g_0}{\partial z} = (z-\zeta)\frac{\mathrm{e}^{\mathrm{j}k_0\sqrt{(z-\zeta)^2+\varrho^2}}\left\{1-\left[(z-\zeta)^2+\varrho^2\right]^{-1/2}\right\}}{4\pi\left[(z-\zeta)^2+\varrho^2\right]}.$$

For $z = 0$, this derivative is an odd function of ζ, so the sum of the two terms satisfies the boundary condition.

Finally, there is no problem with the PDE (2.59) not being satisfied at $\boldsymbol{\xi}_v$ because $\boldsymbol{\xi}_v$ is in the virtual space behind the wall; the solution makes physical sense only in the halfspace $z > 0$.

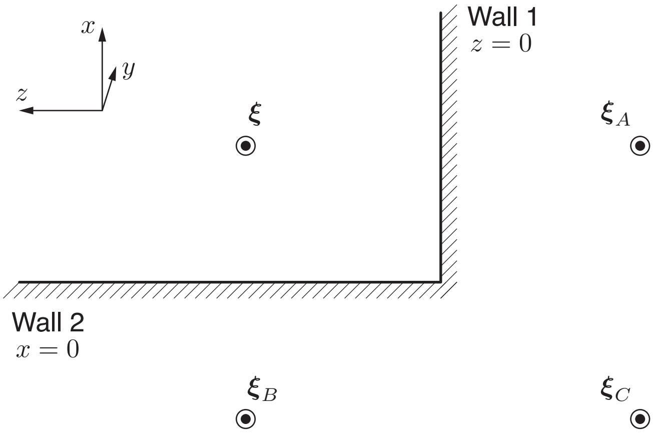


Figure 2.10: Construction of the Green's function for the Helmholtz equation in the intersection of two halfspaces.

Green's Function in a Quarter-Space Before discussing the scope of the IS method, let us work out one more simple example. Consider the hard wall boundary as in Figure 2.10 defining the region $\left\{\boldsymbol{x} = [x, y, z]^\top \,\middle|\, z > 0, x > 0\right\}$. We would like to find the Green's function in this setting by imitating what we did for the halfspace.

We can start by placing a virtual source at $\boldsymbol{\xi}_A$, to make the pressure gradient zero at Wall 1. But this virtual source can only *fix* the solution at Wall 1, so we need to add another one at $\boldsymbol{\xi}_B$ to fix it at Wall 2. Alas, when we add them both, the one at $\boldsymbol{\xi}_A$ changes the gradient at Wall 2, and the one at $\boldsymbol{\xi}_B$ changes the gradient at Wall 1, making them both non-zero.

Intuition suggests that the original source produces pressure at Wall 1 which is matched by the pressure coming from $\boldsymbol{\xi}_A$, resulting in the zero gradient at Wall 1. But $\boldsymbol{\xi}_B$ *helps* the original source from the same side, so A will be too weak to balance it out; similar reasoning holds for the other source. Everything can be balanced out by adding a third virtual source at $\boldsymbol{\xi}_C$ that helps both $\boldsymbol{\xi}_A$ and $\boldsymbol{\xi}_B$ from their sides, making everything click together. But $\boldsymbol{\xi}_C$ can be seen as a second-order image source, a mirror image of a mirror image. That is, noting that $\xi_A = \boldsymbol{\xi} + 2\left\langle \boldsymbol{p}_1 - \boldsymbol{\xi}, \boldsymbol{n}_z \right\rangle \boldsymbol{n}_z$ we can see that $\boldsymbol{\xi}_C = \boldsymbol{\xi}_A + 2\left\langle \boldsymbol{p}_2 - \boldsymbol{\xi}_A, \boldsymbol{n}_x \right\rangle \boldsymbol{n}_x$.

In summary, we have that

$$g(\boldsymbol{x} \,|\, \boldsymbol{\xi}) = g_0(\boldsymbol{x} \,|\, \boldsymbol{\xi}) + g_0(\boldsymbol{x} \,|\, \boldsymbol{\xi}_A) + g_0(\boldsymbol{x} \,|\, \boldsymbol{\xi}_B) + g_0(\boldsymbol{x} \,|\, \boldsymbol{\xi}_C). \tag{2.61}$$

It is straightforward to verify that this function satisfies both the free-space PDE and the boundary conditions.

Scope of the Image Source Method Even though the IS method is used generously in computational acoustics, it is usually justified phenomenologically, by invoking geometrical arguments. When the method is applied inside an enclosure (inside a room), successive reflections of virtual sources generate a sum of fundamental solutions as a candidate for the Green's function. This sum has a finite number of terms only for a limited number of domains, illustrated in Figure 2.11. In any other domain, we obtain an infinite sum.

One should examine the convergence of this sum in some reasonable sense: It may not be convergent, or it may converge conditionally. Keller [102] provided a list of all regions in 2D and

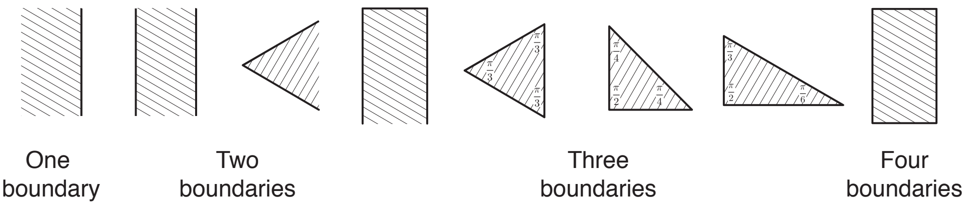


Figure 2.11: Admissible regions for the image method in 2D (exhaustive list).

3D that are admissible for the application of the IS method. He showed that for the method to yield a reasonable solution, all the involved lines (or planes in 3D) must meet at sharp angles that are integer fractions of π.

2.3.3 Need for Geometrical Acoustics

We have seen that the IS method is rigorously valid in a very limited number of domains. Yet, it is used in thousands of research papers to simulate the room acoustics (often in conjunction with much involved stochastic methods for modeling diffuse sound energy distribution, and methods to model edge diffraction). There are highly cited papers extending the IS method to more complex geometries, such as Borish's 1984 "Extension of the Image Method to Arbitrary Polyhedra" [19].

The answer to this conundrum is that the IS method is truly wrong in arbitrary geometries, but it still accurately predicts the behavior or early reflections. To simulate room impulse responses that are perceptually pleasing, and that faithfully model the measured responses, the auralization community developed a host of sophisticated methods to model the energy distribution in the room over time. But the early part is still modeled by the IS model.

As we are primarily interested in early reflections, the IS model is the ideal tool. Using it as an approximation of reality corresponds to *geometrical acoustics*—an acoustical analogy to *geometrical optics*. In geometrical acoustics, we assume that sound can be treated as a collection of rays emanating from, say, a point source. This is of course only an approximation because of a number of wave effects, diffuse reflections, diffractions and so on.

Modeling all these effects exactly is computationally prohibitive, and in fact not necessary. Modern software for simulating the room acoustics produces results that are perceptually indistinguishable from real measurements. At its core are the principles of geometrical acoustics, extended beyond simple reflections to *geometric diffraction theory* [103], and aided by various stochastic methods.

2.3.4 Image Source Model in Rooms

We complete our account of waves and echoes by giving more details about the IS model. We will use this model throughout the thesis, as it allows us to replace simple rooms by point sets.

According to the IS model, we can replace reflections by virtual sources. As illustrated in Figure 2.12, virtual sources are mirror images of the true sources across the corresponding reflecting walls. From the figure, the image $\boldsymbol{s}_i$ of the source $\boldsymbol{s}$ with respect to the ith wall is computed as

$$\boldsymbol{s}_i = \boldsymbol{s} + 2\langle \boldsymbol{p}_i - \boldsymbol{s}, \boldsymbol{n}_i \rangle \boldsymbol{n}_i, \tag{2.62}$$

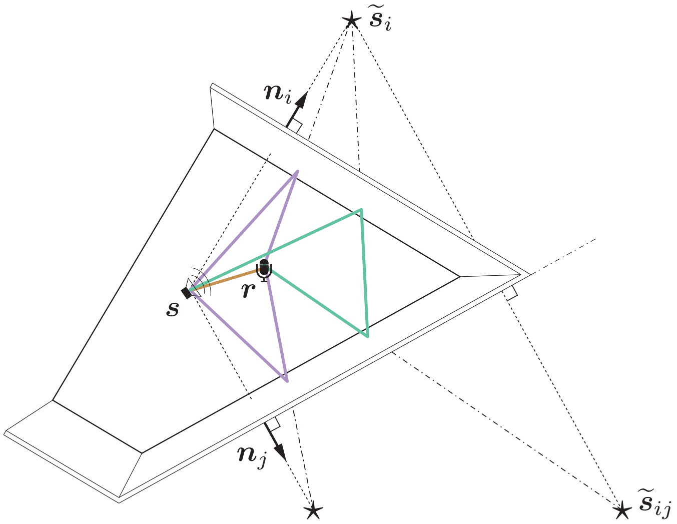


Figure 2.12: Illustration of the image source model for first- and second-order echoes. Vector $\boldsymbol{n}_i$ is the outward-pointing unit normal associated with the ith wall. Stars denote the image sources, and $\widetilde{\boldsymbol{s}}_{ij}$ is the image source corresponding to the second-order echo. Sound rays corresponding to first reflections are shown in purple, and the ray corresponding to the second-order reflection is shown in green.

where $\boldsymbol{n}_i$ is the unit normal, and $\boldsymbol{p}_i$ any point belonging to the ith wall. The time of arrival (TOA) of the echo from the ith wall is $t_i = \|\boldsymbol{s}_i - \boldsymbol{r}\|/c$, where c is the speed of sound. Higher-order image sources corresponding to higher-order echoes are obtained similarly as mirror images of mirror images—we can express them in terms of lower order image sources. For example (again with reference to Figure 2.12) the second-order image source corresponding to walls i and j is given as

$$\boldsymbol{s}_{ij} = \boldsymbol{s}_i + 2\langle \boldsymbol{p}_i - \boldsymbol{s}_i, \boldsymbol{n}_j \rangle \boldsymbol{n}_j. \tag{2.63}$$

Beyond computing the locations of the image sources, we have to take care of their *visibility* and *validity* (details can be found in [19]). As soon as the room has any obtuse angles, some of the image source may not be visible to the receiver, depending on the source-receiver configuration; this indicates the geometric flavor of the method.

2.4 Distances and Euclidean Distance Matrices: From Points to EDMs and Back

We have seen in the previous section that the image source model allows us the model a room by a set of points. When sound is emitted in the room, we can measure the time of arrival of the sound at the microphones, both for the direct sound and for the echoes. Thus what we measure is related to distances; when microphones and sources are synchronized, the TOAs correspond to distances directly.

The principal EDM-related task is to reconstruct the original point set. This task is an inverse problem to the simpler forward problem of finding the EDM given the points. Thus it is desirable to have an analytic expression for the EDM in terms of the point matrix. Beyond convenience, we can expect such an expression to provide interesting structural insights.

Consider a collection of n points in a d-dimensional Euclidean space, ascribed to the columns of matrix $\boldsymbol{X} \in \mathbb{R}^{d\times n}$, $\boldsymbol{X} = [\boldsymbol{x}_1,\ \boldsymbol{x}_2,\ \cdots,\ \boldsymbol{x}_n]$, $\boldsymbol{x}_i \in \mathbb{R}^d$. Then the squared distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ is given as

$$d_{ij} = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2, \tag{2.64}$$

where $\|\cdot\|$ denotes the Euclidean norm. Expanding the norm yields

$$d_{ij} = (\boldsymbol{x}_i - \boldsymbol{x}_j)^\top(\boldsymbol{x}_i - \boldsymbol{x}_j) = \boldsymbol{x}_i^\top \boldsymbol{x}_i - 2\boldsymbol{x}_i^\top \boldsymbol{x}_j + \boldsymbol{x}_j^\top \boldsymbol{x}_j. \tag{2.65}$$

From here, we can read out the matrix equation for $\boldsymbol{D} = [d_{ij}]$,

$$\operatorname{EDM}(\boldsymbol{X}) \overset{\text{def}}{=} \boldsymbol{1}\operatorname{diag}(\boldsymbol{X}^\top\boldsymbol{X})^\top - 2\boldsymbol{X}^\top\boldsymbol{X} + \operatorname{diag}(\boldsymbol{X}^\top\boldsymbol{X})\boldsymbol{1}^\top, \tag{2.66}$$

where $\boldsymbol{1}$ denotes the column vector of all ones and $\operatorname{diag}(\boldsymbol{A})$ is a column vector of the diagonal entries of $\boldsymbol{A}$. We see that $\operatorname{EDM}(\boldsymbol{X})$ is in fact a function of $\boldsymbol{X}^\top\boldsymbol{X}$. For later reference, it is convenient to define an operator $\mathcal{K}(\boldsymbol{G})$ similar to $\operatorname{EDM}(\boldsymbol{X})$, that operates directly on the Gram matrix $\boldsymbol{G} = \boldsymbol{X}^\top\boldsymbol{X}$,

$$\mathcal{K}(\boldsymbol{G}) \overset{\text{def}}{=} \operatorname{diag}(\boldsymbol{G})\boldsymbol{1}^\top - 2\boldsymbol{G} + \boldsymbol{1}\operatorname{diag}(\boldsymbol{G})^\top. \tag{2.67}$$

The EDM assembly formula (2.66) or (2.67) reveals an important property: Because the rank of $\boldsymbol{X}$ is at most d (it has d rows), then the rank of $\boldsymbol{X}^\top\boldsymbol{X}$ is also at most d. The remaining two summands in (2.66) have rank one. By rank inequalities, the rank of a sum of matrices cannot exceed the sum of the ranks of the summands. With this observation, we proved one of the most notable facts about EDMs:

Theorem 2.1 (*Rank of EDMs*)

Rank of an EDM corresponding to points in $\mathbb{R}^d$ is at most $d + 2$.

This is a powerful theorem: it states that the rank of an EDM is independent of the number of points that generate it. In many applications, d is three or less, while n can be in the thousands. According to Theorem 2.1, rank of such practical matrices is at most five. The proof of this theorem is simple, but to appreciate that the property is not obvious, you may try to compute the rank of the matrix of non-squared distances.

What really matters in Theorem 2.1 is the affine dimension of the point set—the dimension of the smallest affine subspace that contains the points, denoted by $\operatorname{affdim}(\boldsymbol{X})$. For example, if the points lie on a plane (but not on a line or a circle) in $\mathbb{R}^3$, rank of the corresponding EDM is four, not five. This will be clear from a different perspective in the next subsection, as any affine subspace is just a translation of a linear subspace. An illustration for a 1D subspace of $\mathbb{R}^2$ is provided in Figure 2.13: Subtracting *any* point in the affine subspace from all its points translates it to the parallel linear subspace that contains the zero vector.

2.4.1 Essential Uniqueness

When solving an inverse problem, we need to understand what is recoverable and what is forever lost in the forward problem. Representing sets of points by distances usually increases the size of the representation. For most interesting n and d, the number of pairwise distances is larger than the size of the coordinate description, $\binom{n}{2} > nd$, so an EDM holds more scalars than the list of point coordinates. Nevertheless, some information is lost in this encoding, namely the

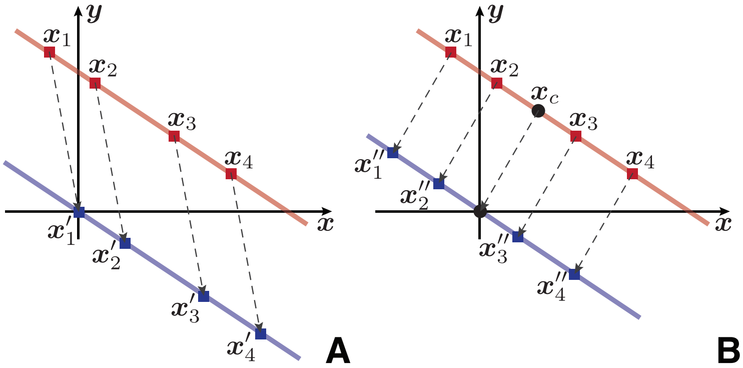


Figure 2.13: Illustration of the relationship between an affine subspace and its parallel linear subspace. The points $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_4]$ live in an affine subspace—a line in $\mathbb{R}^2$ that does not contain the origin. In (A), the vector $\boldsymbol{x}_1$ is subtracted from all the points, and the new point list is $\boldsymbol{X}' = [\boldsymbol{0}, \boldsymbol{x}_2 - \boldsymbol{x}_1, \boldsymbol{x}_3 - \boldsymbol{x}_1, \boldsymbol{x}_4 - \boldsymbol{x}_1]$. While the columns of $\boldsymbol{X}$ span $\mathbb{R}^2$, the columns of $\boldsymbol{X}'$ only span the 1D subspace of $\mathbb{R}^2$—the line through the origin. In (B), we subtract a different vector from all points: the centroid $\frac{1}{4}\boldsymbol{X}\boldsymbol{1}$. The translated vectors $\boldsymbol{X}'' = [\boldsymbol{x}_1'', \ldots, \boldsymbol{x}_4'']$ again span the same 1D subspace.

information about the absolute position and orientation of the point set. Intuitively, it is clear that rigid transformations (including reflections) do not change distances between the fixed points in a point set. This intuitive fact is easily deduced from the EDM assembly formula (2.66). We have seen in (2.66) and (2.67) that $\mathrm{EDM}(\boldsymbol{X})$ is in fact a function of the Gram matrix $\boldsymbol{X}^\top \boldsymbol{X}$.

This makes it easy to show algebraically that rotations and reflections do not alter the distances. Any rotation/reflection can be represented by an orthogonal matrix $\boldsymbol{Q} \in \mathbb{R}^{d \times d}$ acting on the points $\boldsymbol{x}_i$. Thus for the rotated point set $\boldsymbol{X}_r = \boldsymbol{Q}\boldsymbol{X}$ we can write

$$\boldsymbol{X}_r^\top \boldsymbol{X}_r = (\boldsymbol{Q}\boldsymbol{X})^\top (\boldsymbol{Q}\boldsymbol{X}) = \boldsymbol{X}^\top \boldsymbol{Q}^\top \boldsymbol{Q}\boldsymbol{X} = \boldsymbol{X}^\top \boldsymbol{X}, \tag{2.68}$$

where we invoked the orthogonality of the rotation/reflection matrix, $\boldsymbol{Q}^\top \boldsymbol{Q} = \boldsymbol{I}$.

Translation by a vector $\boldsymbol{b} \in \mathbb{R}^d$ can be expressed as

$$\boldsymbol{X}_t = \boldsymbol{X} + \boldsymbol{b}\boldsymbol{1}^\top. \tag{2.69}$$

Using $\mathrm{diag}(\boldsymbol{X}_t^\top \boldsymbol{X}_t) = \mathrm{diag}(\boldsymbol{X}^\top \boldsymbol{X}) + 2\boldsymbol{X}^\top \boldsymbol{b} + \|b\|^2 \boldsymbol{1}$, one can directly verify that this transformation leaves (2.66) intact. In summary,

$$\mathrm{EDM}(\boldsymbol{Q}\boldsymbol{X}) = \mathrm{EDM}(\boldsymbol{X} + \boldsymbol{b}\boldsymbol{1}^\top) = \mathrm{EDM}(\boldsymbol{X}). \tag{2.70}$$

The consequence of this invariance is that we will never be able to reconstruct the absolute orientation of the point set using only the distances, and the corresponding degrees of freedom will be chosen freely. Different reconstruction procedures will lead to different realizations of the point set, all of them being rigid transformations of each other. Figure 2.14 illustrates a point set under a rigid transformation. It is clear that the distances between the points are the same for all three shapes.

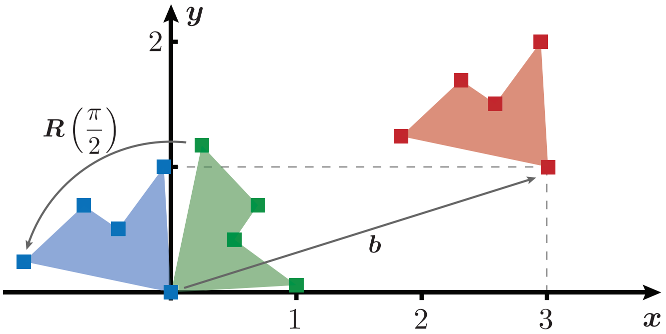


Figure 2.14: Illustration of a rigid transformation in 2D. Here the points set is transformed as $\boldsymbol{R}\boldsymbol{X} + \boldsymbol{b}\boldsymbol{1}^\top$, but the corresponding EDM does not change. Rotation matrix $\boldsymbol{R} = \left[\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right]$, corresponds to a counterclockwise rotation of 90°. The translation vector is $\boldsymbol{b} = [3,\ 1]^\top$. The shape is drawn for visual reference.

2.4.2 Reconstructing the Point Set From Distances

The EDM equation (2.66) hints at a procedure to compute the point set starting from the distance matrix. Consider the following choice: let the first point $\boldsymbol{x}_1$ be at the origin. Then the first column of $\boldsymbol{D}$ contains the squared norms of the point vectors,

$$d_{i1} = \|\boldsymbol{x}_i - \boldsymbol{x}_1\|^2 = \|\boldsymbol{x}_i - \boldsymbol{0}\|^2 = \|\boldsymbol{x}_i\|^2. \tag{2.71}$$

Consequently, we can immediately construct the term $\boldsymbol{1}\operatorname{diag}(\boldsymbol{X}^\top \boldsymbol{X})$ and its transpose in (2.66), as the diagonal of $\boldsymbol{X}^\top \boldsymbol{X}$ contains exactly the norms squared $\|\boldsymbol{x}_i\|^2$. Concretely,

$$\boldsymbol{1}\operatorname{diag}(\boldsymbol{X}^\top \boldsymbol{X}) = \boldsymbol{1}\,\boldsymbol{d}_1^\top, \tag{2.72}$$

where $\boldsymbol{d}_1 = \boldsymbol{D}\boldsymbol{e}_1$ is the first column of $\boldsymbol{D}$. We thus obtain the Gram matrix from (2.66) as

$$\boldsymbol{G} = \boldsymbol{X}^\top \boldsymbol{X} = -\frac{1}{2}(\boldsymbol{D} - \boldsymbol{1}\,\boldsymbol{d}_1^\top - \boldsymbol{d}_1\boldsymbol{1}^\top). \tag{2.73}$$

The point set can be found by an eigenvalue decomposition (EVD), $\boldsymbol{G} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$, where $\boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ with all eigenvalues λ_i non-negative, and $\boldsymbol{U}$ orthonormal, as $\boldsymbol{G}$ is a symmetric positive semidefinite matrix. Throughout the chapter we assume that the eigenvalues are sorted in the order of decreasing magnitude, $|\lambda_1| \geqslant |\lambda_2| \geqslant \cdots \geqslant |\lambda_n|$. We can now set $\widehat{\boldsymbol{X}} \overset{\text{def}}{=} \left[\operatorname{diag}\left(\sqrt{\lambda_1},\ \ldots,\ \sqrt{\lambda_d}\right),\ \boldsymbol{0}_{d\times(n-d)}\right]\boldsymbol{U}^\top$. Note that we could have simply taken $\boldsymbol{\Lambda}^{1/2}\boldsymbol{U}^\top$ as the reconstructed point set, but if the Gram matrix really describes a d-dimensional point set, the trailing eigenvalues will be zeros, so we choose to truncate the corresponding rows.

It is straightforward to verify that the reconstructed point set $\widehat{\boldsymbol{X}}$ generates the original EDM, $\boldsymbol{D} = \operatorname{EDM}(\boldsymbol{X})$; as we have learned, $\widehat{\boldsymbol{X}}$ and $\boldsymbol{X}$ are related by a rigid transformation. The described procedure is called the *classical MDS*, with a particular choice of the coordinate system: $\boldsymbol{x}_1$ is fixed at the origin.

In (2.73) we subtract a structured rank-2 matrix $(\boldsymbol{1}\,\boldsymbol{d}_1^\top + \boldsymbol{d}_1\boldsymbol{1}^\top)$ from $\boldsymbol{D}$. A more systematic approach to the classical MDS is to use a generalization of (2.73) by Gower [83]. Any such subtraction that makes the right hand side of (2.73) positive semidefinite (PSD), *i.e.*, that makes $\boldsymbol{G}$ a Gram matrix, can also be modeled by multiplying $\boldsymbol{D}$ from both sides by a particular matrix. This is substantiated in the following result.

Algorithm 2.1 Classical MDS

```
function ClassicalMDS($\boldsymbol{D}, d$)
    $\boldsymbol{J} \leftarrow \boldsymbol{I} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^\top$        ▷ Geometric centering matrix
    $\boldsymbol{G} \leftarrow -\frac{1}{2}\boldsymbol{J}\boldsymbol{D}\boldsymbol{J}$        ▷ Compute the Gram matrix
    $\boldsymbol{U}, [\lambda_i]_{i=1}^n \leftarrow \mathrm{EVD}(\boldsymbol{G})$
    return $\left[\mathrm{diag}\left(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_d}\right), \boldsymbol{0}_{d\times(n-d)}\right]\boldsymbol{U}^\top$
end function
```

Theorem 2.2 (*Gower [83]*)

$\boldsymbol{D}$ is an EDM if and only if

$$-\frac{1}{2}(\boldsymbol{I} - \boldsymbol{1}\boldsymbol{s}^\top)\boldsymbol{D}(\boldsymbol{I} - \boldsymbol{s}\boldsymbol{1}^\top) \tag{2.74}$$

is PSD for any $\boldsymbol{s}$ such that $\boldsymbol{s}^\top\boldsymbol{1} = 1$ and $\boldsymbol{s}^\top\boldsymbol{D} \neq \boldsymbol{0}$.

In fact, if (2.74) is PSD for one such $\boldsymbol{s}$, then it is PSD for all of them. In particular, define the *geometric centering matrix* as

$$\boldsymbol{J} \overset{\text{def}}{=} \boldsymbol{I} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^\top. \tag{2.75}$$

Then $-\frac{1}{2}\boldsymbol{J}\boldsymbol{D}\boldsymbol{J}$ being positive semidefinite is equivalent to $\boldsymbol{D}$ being an EDM. Different choices of $\boldsymbol{s}$ correspond to different translations of the point set.

The classical MDS algorithm with the geometric centering matrix is spelled out in Algorithm 2.1. Whereas so far we have assumed that the distance measurements are noiseless, Algorithm 2.1 can handle noisy distances too, as it discards all but the d largest eigenvalues.

It is straightforward to verify that (2.73) corresponds to $\boldsymbol{s} = \boldsymbol{e}_1$. Think about what this means in terms of the point set: $\boldsymbol{X}\boldsymbol{e}_1$ selects the first point in the list, $\boldsymbol{x}_1$. Then $\boldsymbol{X}_0 = \boldsymbol{X}(\boldsymbol{I} - \boldsymbol{e}_1\boldsymbol{1}^\top)$ translates the points so that $\boldsymbol{x}_1$ is translated to the origin. Multiplying the definition (2.66) from the right by $(\boldsymbol{I} - \boldsymbol{e}_1\boldsymbol{1}^\top)$ and from the left by $(\boldsymbol{I} - \boldsymbol{1}\boldsymbol{e}_1^\top)$ will annihilate the two rank-1 matrices, $\mathrm{diag}(\boldsymbol{G})\boldsymbol{1}^\top$ and $\boldsymbol{1}\,\mathrm{diag}(\boldsymbol{G})^\top$. We see that the remaining term has the form $-2\boldsymbol{X}_0^\top\boldsymbol{X}_0$, and the reconstructed point set will have the first point at the origin!

On the other hand, setting $\boldsymbol{s} = \frac{1}{n}\boldsymbol{1}$ places the centroid of the point set at the origin of the coordinate system. For this reason, the matrix $\boldsymbol{J} = \boldsymbol{I} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^\top$ is called the *centering matrix*. To better understand why, consider how we normally center a set of points given in $\boldsymbol{X}$.

First, we compute the centroid as the mean of all the points

$$\boldsymbol{x}_c = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i = \frac{1}{n}\boldsymbol{X}\boldsymbol{1}. \tag{2.76}$$

Second, we subtract this vector from all the points in the set,

$$\boldsymbol{X}_c = \boldsymbol{X} - \boldsymbol{x}_c\boldsymbol{1}^\top = \boldsymbol{X} - \frac{1}{n}\boldsymbol{X}\boldsymbol{1}\boldsymbol{1}^\top = \boldsymbol{X}(\boldsymbol{I} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^\top). \tag{2.77}$$

In complete analogy with the reasoning for $\boldsymbol{s} = \boldsymbol{e}_1$, we can see that the reconstructed point set will be centered at the origin.

2.4.3 Orthogonal Procrustes Problem

Since the absolute position and orientation of the points are lost when going over to distances, we need a method to align the reconstructed point set with a set of *anchors*—points whose coordinates are fixed and known.

This can be achieved in two steps, sometimes called Procrustes analysis. Ascribe the anchors to the columns of $\boldsymbol{Y}$, and suppose that we want to align the point set $\boldsymbol{X}$ with the columns of $\boldsymbol{Y}$. Let $\boldsymbol{X}_a$ denote the submatrix (a selection of columns) of $\boldsymbol{X}$ that should be aligned with the anchors. We note that the number of anchors—columns in $\boldsymbol{X}_a$—is typically small compared with the total number of points—columns in $\boldsymbol{X}$.

In the first step, we remove the means $\boldsymbol{y}_c$ and $\boldsymbol{x}_{a,c}$ from matrices $\boldsymbol{Y}$ and $\boldsymbol{X}_a$, obtaining the matrices $\overline{\boldsymbol{Y}}$ and $\overline{\boldsymbol{X}}_a$. In the second step, termed orthogonal Procrustes analysis, we are searching for the rotation and reflection that best maps $\overline{\boldsymbol{X}}_a$ onto $\overline{\boldsymbol{Y}}$,

$$\boldsymbol{R} = \underset{\boldsymbol{Q}:\boldsymbol{Q}\boldsymbol{Q}^\top=\boldsymbol{I}}{\arg\min} \|\boldsymbol{Q}\overline{\boldsymbol{X}}_a - \overline{\boldsymbol{Y}}\|_F^2. \tag{2.78}$$

The Frobenius norm $\|\cdot\|_F$ is simply the ℓ^2-norm of the matrix entries, $\|\boldsymbol{A}\|_F^2 \overset{\text{def}}{=} \sum a_{ij}^2 = \operatorname{trace}(\boldsymbol{A}^\top\boldsymbol{A})$.

The solution to (2.78)—found by Schönemann in his PhD thesis [185]—is given by the singular value decomposition (SVD). Let $\overline{\boldsymbol{X}}_a\overline{\boldsymbol{Y}}^\top = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$; then we can continue computing (2.78) as follows

$$\begin{aligned}\boldsymbol{R} &= \underset{\boldsymbol{Q}:\boldsymbol{Q}\boldsymbol{Q}^\top=\boldsymbol{I}}{\arg\min} \|\boldsymbol{Q}\overline{\boldsymbol{X}}_a\|_F^2 + \|\overline{\boldsymbol{Y}}\|_F^2 - \operatorname{trace}(\overline{\boldsymbol{Y}}^\top\boldsymbol{Q}\overline{\boldsymbol{X}}_a)\\ &= \underset{\widetilde{\boldsymbol{Q}}:\widetilde{\boldsymbol{Q}}\widetilde{\boldsymbol{Q}}^\top=\boldsymbol{I}}{\arg\max} \operatorname{trace}(\widetilde{\boldsymbol{Q}}\,\boldsymbol{\Sigma}),\end{aligned} \tag{2.79}$$

where $\widetilde{\boldsymbol{Q}} \overset{\text{def}}{=} \boldsymbol{V}^\top\boldsymbol{Q}\boldsymbol{U}$, and we used the orthogonal invariance of the Frobenius norm and the cyclic invariance of the trace. The last trace expression in (2.79) is equal to $\sum_{i=1}^n \sigma_i\widetilde{q}_{ii}$. Noting that $\widetilde{\boldsymbol{Q}}$ is also an orthogonal matrix, its diagonal entries cannot exceed 1. Therefore, the maximum is achieved when $\widetilde{q}_{ii} = 1$ for all i, meaning that the optimal $\widetilde{\boldsymbol{Q}}$ is an identity matrix. It follows that $\boldsymbol{R} = \boldsymbol{V}\boldsymbol{U}^\top$.

Once the optimal rigid transformation has been found, the alignment can be applied to the entire point set as

$$\boldsymbol{R}(\boldsymbol{X} - \boldsymbol{x}_{a,c}\boldsymbol{1}^\top) + \boldsymbol{y}_c\boldsymbol{1}^\top. \tag{2.80}$$

2.4.4 Counting the Degrees of Freedom

It is interesting to count how many degrees of freedom there are in different EDM related objects. Clearly, for n points in $\mathbb{R}^d$ we have

$$\#_{\boldsymbol{X}} = n \times d \tag{2.81}$$

degrees of freedom: If we describe the point set by the list of coordinates, the size of the description matches the number of degrees of freedom. Going from the points to the EDM (usually) increases the description size to $\frac{1}{2}n(n-1)$, as the EDM lists the distances between all the pairs of points. By Theorem 2.1 we know that the EDM has rank at most $d+2$.

Let us imagine for a moment that we do not know any other EDM-specific properties of our matrix, except that it is symmetric, positive, zero-diagonal (or *hollow*), and that it has rank $d+2$. The purpose of this exercise is to count the degrees of freedom associated with such a matrix, and to see if their number matches the intrinsic number of the degrees of freedom of the point set, $\#_{\boldsymbol{X}}$. If it did, then these properties would completely characterize an EDM. We can already anticipate from Theorem 2.2 that we need more properties: a certain matrix related to the EDM—as given in (2.74)—must be PSD. Still, we want to see how many degrees of freedom we miss.

We can do the counting by looking at the EVD of a symmetric matrix, $\boldsymbol{D} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$. The diagonal matrix $\boldsymbol{\Lambda}$ is specified by $d+2$ degrees of freedom, because $\boldsymbol{D}$ has rank $d+2$. The first eigenvector of length n takes up $n-1$ degrees of freedom due to the normalization; the second one takes up $n-2$, as it is in addition orthogonal to the first one; for the last eigenvector, number $(d+2)$, we need $n-(d+2)$ degrees of freedom. We do not need to count the other eigenvectors, because they correspond to zero eigenvalues. The total number is then

$$\begin{aligned}\#_{\text{DOF}} &= \underbrace{(d+2)}_{\text{Eigenvalues}} + \underbrace{(n-1)+\cdots+[n-(d+2)]}_{\text{Eigenvectors}} - \underbrace{n}_{\text{Hollowness}} \\ &= n \times (d+1) - \frac{(d+1)\times(d+2)}{2}\end{aligned}$$

For large n and fixed d, it follows that

$$\frac{\#_{\text{DOF}}}{\#_{\boldsymbol{X}}} \sim \frac{d+1}{d}. \tag{2.82}$$

Therefore, even though the rank property is useful and we will show efficient algorithms that exploit it, it is still not a *tight* property (symmetry and hollowness included). For $d=3$, the ratio (2.82) is $\frac{4}{3}$, so, loosely speaking, the rank property has 30% determining scalars too many, which we need to set consistently. Put differently, we need 30% more data in order to exploit the rank property than we need to exploit the full EDM structure. Again loosely phrased, we can assert that for the same amount of data, the algorithms perform at least ≈30% worse if we only exploit the rank property, without *EDMness.*

The one-third gap accounts for various geometrical constraints that must be satisfied. The redundancy in the EDM representation is what makes denoising and completion algorithms possible, and thinking in terms of degrees of freedom gives us a fundamental understanding of what is achievable. Interestingly, the above discussion suggests that for large n and large $d = o(n)$, little is lost by only considering rank.

Finally, in the above discussion, for the sake of simplicity, we ignored the degrees of freedom related to the absolute orientation. These degrees of freedom, not present in the EDM, do not affect the large-n behavior.

2.5 EDMs as a Practical Tool

We rarely have a perfect EDM. Not only are the entries of the measured matrix plagued by errors, but often we can measure just a subset. There are various sources of error in distance measurements: we already know that in NMR spectroscopy, instead of exact distances we get intervals. Measuring distance using received powers or TOAs is subject to noise, sampling errors and model mismatch.

Missing entries arise because of the limited radio range, or because of the nature of the spectrometer. Sometimes the nodes in the problem at hand are asymmetric by definition; in microphone calibration we have two types: microphones and calibration sources. This results in a particular block structure of the missing entries (we will come back to this in Chapter 4, but you can fast-forward to Figure 4.11 for an illustration).

It is convenient to have a single statement for both EDM approximation and EDM completion, as the algorithms described in this section handle them at once.

Problem 2.1

> *Let* $\boldsymbol{D} = \mathrm{EDM}(\boldsymbol{X})$*. We are given a noisy observation of the distances between* $p \leqslant \binom{n}{2}$ *pairs of points from* $\boldsymbol{X}$*. That is, we have a noisy measurement of* $2p$ *entries in* $\boldsymbol{D}$*,*
>
> $$\widetilde{d}_{ij} = d_{ij} + \epsilon_{ij}, \tag{2.83}$$
>
> *for* $(i, j) \in E$*, where* E *is some index set, and* ϵ_{ij} *absorbs all errors. The goal is to reconstruct the point set* $\widehat{\boldsymbol{X}}$ *in the given embedding dimension, so that the entries of* $\mathrm{EDM}(\widehat{\boldsymbol{X}})$ *are close in some metric to the observed entries* $\widetilde{d}_{ij}$*.*

To concisely write down completion problems, we define the mask matrix $\boldsymbol{W}$ as follows,

$$w_{ij} \overset{\mathrm{def}}{=} \begin{cases} 1, & (i, j) \in E \\ 0, & \text{otherwise.} \end{cases} \tag{2.84}$$

This matrix then selects elements of an EDM through a Hadamard (entrywise) product. For example, to compute the norm of the difference between the observed entries in $\boldsymbol{A}$ and $\boldsymbol{B}$, we write $\| \boldsymbol{W} \circ (\boldsymbol{A} - \boldsymbol{B}) \|$. Furthermore, we define the indexing $\boldsymbol{A}_{\boldsymbol{W}}$ to mean the restriction of $\boldsymbol{A}$ to those entries where $\boldsymbol{W}$ is non-zero. The meaning of $\boldsymbol{B}_{\boldsymbol{W}} \leftarrow \boldsymbol{A}_{\boldsymbol{W}}$ is that we assign the observed part of $\boldsymbol{A}$ to the observed part of $\boldsymbol{B}$.

2.5.1 Exploiting the Rank Property

Perhaps the most notable fact about EDMs is the rank property established in Theorem 2.1: The rank of an EDM for points living in $\mathbb{R}^d$ is at most $d + 2$. This leads to conceptually simple algorithms for EDM completion and denoising. Interestingly, these algorithms exploit only the rank of the EDM. There is no explicit Euclidean geometry involved, at least not before reconstructing the point set.

We have two pieces of information: a subset of potentially noisy distances, and the desired embedding dimension of the point configuration. The latter implies the rank property of the EDM that we aim to exploit. We may try to alternate between enforcing these two properties, and hope that the algorithm produces a sequence of matrices that converges to an EDM. If it does, we have a solution. Alternatively, it may happen that we converge to a matrix with the correct rank that is not an EDM, or that the algorithm never converges. The pseudocode is listed in Algorithm 2.2.

A different, more powerful approach is to leverage algorithms for low rank matrix completion developed by the compressed sensing community. For example, OptSpace [104] is an algorithm for

Algorithm 2.2 Alternating Rank-Based EDM Completion

```
function RankCompleteEDM($\boldsymbol{W}, \widetilde{\boldsymbol{D}}, d$)
    $\boldsymbol{D}_{\boldsymbol{W}} \leftarrow \widetilde{\boldsymbol{D}}_{\boldsymbol{W}}$                ▷ Initialize observed entries
    $\boldsymbol{D}_{\boldsymbol{1}\boldsymbol{1}^\top - \boldsymbol{W}} \leftarrow \mu$                ▷ Initialize unobserved entries
    repeat
        $\boldsymbol{D} \leftarrow$ EVThreshold($\boldsymbol{D}$, $d + 2$)
        $\boldsymbol{D}_{\boldsymbol{W}} \leftarrow \widetilde{\boldsymbol{D}}_{\boldsymbol{W}}$                ▷ Enforce known entries
        $\boldsymbol{D}_{\boldsymbol{I}} \leftarrow \boldsymbol{0}$                ▷ Set the diagonal to zero
        $\boldsymbol{D} \leftarrow (\boldsymbol{D})_+$                ▷ Zero the negative entries
    until Convergence or MaxIter
    return $\boldsymbol{D}$
end function

function EVThreshold($\boldsymbol{D}, r$)
    $\boldsymbol{U}, [\lambda_i]_{i=1}^n \leftarrow$ EVD($\boldsymbol{D}$)
    $\boldsymbol{\Sigma} \leftarrow \mathrm{diag}\big(\lambda_1, \ldots, \lambda_r, \underbrace{0, \ldots, 0}_{n-r \text{ times}}\big)$
    $\boldsymbol{D} \leftarrow \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}^\top$
    return $\boldsymbol{D}$
end function
```

recovering a low-rank matrix from noisy, incomplete data. Let us take a look at how OptSpace works. Denote by $\boldsymbol{M} \in \mathbb{R}^{m \times n}$ the rank-r matrix that we seek to recover, by $\boldsymbol{Z} \in \mathbb{R}^{m \times n}$ the measurement noise, and by $\boldsymbol{W} \in \mathbb{R}^{m \times n}$ the mask corresponding to the measured entries; for simplicity we choose $m \leqslant n$. The measured noisy and incomplete matrix is then given as

$$\widetilde{\boldsymbol{M}} = \boldsymbol{W} \circ (\boldsymbol{M} + \boldsymbol{Z}). \tag{2.85}$$

Effectively, this sets the missing (non-observed) entries of the matrix to zero. OptSpace aims to minimize the following cost function,

$$F(\boldsymbol{A}, \boldsymbol{S}, \boldsymbol{B}) \overset{\text{def}}{=} \frac{1}{2} \| \boldsymbol{W} \circ (\widetilde{\boldsymbol{M}} - \boldsymbol{A}\boldsymbol{S}\boldsymbol{B}^\top) \|_F^2, \tag{2.86}$$

where $\boldsymbol{S} \in \mathbb{R}^{r \times r}$, $\boldsymbol{A} \in \mathbb{R}^{m \times r}$, and $\boldsymbol{B} \in \mathbb{R}^{n \times r}$ such that $\boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{B}^\top \boldsymbol{B} = \boldsymbol{I}$. Note that $\boldsymbol{S}$ need not be diagonal.

The cost function (2.86) is not convex, and minimizing it is *a priori* difficult [105] due to many local minima. Nevertheless, Keshavan, Montanari and Oh [104] show that using the gradient descent method to solve (2.86) yields the global optimum with high probability, provided that the descent is correctly initialized.

Let $\widetilde{\boldsymbol{M}} = \sum_{i=1}^m \sigma_i \boldsymbol{a}_i \boldsymbol{b}_i^\top$ be the SVD of $\widetilde{\boldsymbol{M}}$. Then we define the scaled rank-r projection of $\widetilde{\boldsymbol{M}}$ as $\widetilde{\boldsymbol{M}}_r \overset{\text{def}}{=} \alpha^{-1} \sum_{i=1}^r \sigma_i \boldsymbol{a}_i \boldsymbol{b}_i^\top$. The fraction of observed entries is denoted by α, so that the scaling factor compensates the smaller *average* magnitude of the entries in $\widetilde{\boldsymbol{M}}$ in comparison with $\boldsymbol{M}$. The SVD of $\widetilde{\boldsymbol{M}}_r$ is then used to initialize the gradient descent, as detailed in Algorithm 2.3.

Two additional remarks are due in the description of OptSpace. First, it can be shown that the performance is improved by zeroing the *over-represented* rows and columns. A row (*resp.* column) is over-represented if it contains more than twice the average number of observed entries per row (*resp.* column). These heavy rows and columns bias the corresponding singular vectors

Algorithm 2.3 OptSpace [104]

```
1: function OptSpace($\widetilde{\boldsymbol{M}}, r$)
2:     $\widetilde{\boldsymbol{M}} \leftarrow \mathrm{Trim}(\widetilde{\boldsymbol{M}})$
3:     $\widetilde{\boldsymbol{A}}, \widetilde{\boldsymbol{\Sigma}}, \widetilde{\boldsymbol{B}} \leftarrow \mathrm{SVD}(\alpha^{-1}\widetilde{\boldsymbol{M}})$
4:     $\boldsymbol{A}_0 \leftarrow$ First $r$ columns of $\widetilde{\boldsymbol{A}}$
5:     $\boldsymbol{B}_0 \leftarrow$ First $r$ columns of $\widetilde{\boldsymbol{B}}$
6:     $\boldsymbol{S}_0 \leftarrow \arg\min_{\boldsymbol{S} \in \mathbb{R}^{r \times r}} F(\boldsymbol{A}_0, \boldsymbol{S}, \boldsymbol{B}_0)$        ▷ Eq. (2.86)
7:     $\boldsymbol{A}, \boldsymbol{B} \leftarrow \arg\min_{\boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{B}^\top \boldsymbol{B} = \boldsymbol{I}} F(\boldsymbol{A}, \boldsymbol{S}_0, \boldsymbol{B})$        ▷ See the note below
8:     return $\boldsymbol{A}\boldsymbol{S}_0\boldsymbol{B}^\top$
9: end function
▷ Line 7: gradient descent starting at $\boldsymbol{A}_0, \boldsymbol{B}_0$
```

and singular values, so (perhaps surprisingly) it is better to throw them away. We call this step "Trim" in Algorithm 2.3.

Second, the minimization of (2.86) does not have to be performed for all variables at once. In [104], the authors first solve the easier, convex minimization for $\boldsymbol{S}$, and then with the optimizer $\boldsymbol{S}$ fixed, they find the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ using the gradient descent. These steps correspond to lines 6 and 7 of Algorithm 2.3.

2.5.2 Multidimensional Scaling

Multidimensional scaling refers to a group of techniques that, given a set of noisy distances, find the best fitting point conformation. It was originally proposed in psychometrics [114, 206] to visualize the (dis-)similarities between objects. Initially, MDS was defined as the *problem* of representing distance data, but now the term is commonly used to refer to *methods* for solving the problem [18].

Various cost functions were proposed for solving MDS. In Section 2.4.2, we already encountered one method: the classical MDS. This method minimizes the Frobenius norm of the difference between the input matrix and the Gram matrix of the points in the target embedding dimension.

The Gram matrix contains inner products, but it is better to directly work with the distances. A typical cost function represents the dissimilarity of the observed distances and the distances between the estimated point locations. An essential observation is that the feasible set for these optimizations is not convex (EDMs with embedding dimensions smaller than $n-1$ lie on the boundary of a cone [44], which is a non-convex set).

A popular dissimilarity measure is *raw stress* [114], defined as the value of

$$\underset{\boldsymbol{X} \in \mathbb{R}^{d \times n}}{\text{minimize}} \quad \sum_{(i,j) \in E} \left(\sqrt{\mathrm{EDM}(\boldsymbol{X})_{ij}} - \sqrt{\tilde{d}_{ij}} \right)^2, \tag{2.87}$$

where E defines the set of revealed elements of the distance matrix $\boldsymbol{D}$. The objective function can be concisely written as $\left\| \boldsymbol{W} \circ \left(\sqrt{\mathrm{EDM}(\boldsymbol{X})} - \sqrt{\widetilde{\boldsymbol{D}}} \right) \right\|_F^2$; a drawback of this cost function is that it is not globally differentiable.

Another well-known cost function, first studied by Takane, Young and De Leeuw [198], is *s-stress*,

$$\underset{\boldsymbol{X}\in\mathbb{R}^{d\times n}}{\text{minimize}} \quad \sum_{(i,j)\in E} \Big(\text{EDM}(\boldsymbol{X})_{ij} - \widetilde{d}_{ij}\Big)^2 . \tag{2.88}$$

Again, we write the objective concisely as $\big\| \boldsymbol{W} \circ \big(\text{EDM}(\boldsymbol{X}) - \widetilde{\boldsymbol{D}}\big)\big\|_F^2$. Conveniently, the s-stress objective is everywhere differentiable, but at a disadvantage that it puts more weight on errors in larger distances. Gaffke and Mathar [76] propose an algorithm to find the global minimum of the s-stress function for embedding dimension $d = n - 1$. EDMs with this embedding dimension exceptionally constitute a convex set [44], but we are typically interested in embedding dimensions much smaller than n. The s-stress minimization in (2.88) is not convex for $d < n - 1$. It was analytically shown to have saddle points [162], but interestingly, no analytical non-global minimizer has been found [162].

Browne proposed a method for computing s-stress based on Newton-Raphson root finding [26]. Glunt reports that the method by Browne converges to the global minimum of (2.88) in 90% of the test cases in his dataset[2] [78].

The cost function in (2.88) is separable across points i and across coordinates k, which is convenient for distributed implementations. Parhizkar [162] proposed an alternating coordinate descent method that leverages this separability, by updating a single coordinate of a particular point at a time. The s-stress function restricted to the kth coordinate of the ith point is a fourth-order polynomial,

$$f(x; \boldsymbol{\alpha}^{(i,k)}) = \sum_{\ell=0}^{4} \alpha_\ell^{(i,k)} x^\ell, \tag{2.89}$$

where $\boldsymbol{\alpha}^{(i,k)}$ lists the polynomial coefficients for ith point and kth coordinate. For example, $\alpha_0^{(i,k)} = 4\sum_j w_{ij}$, that is, four times the number of points connected to point i. Expressions for the remaining coefficients are given in [162]; in the pseudocode (Algorithm 2.4), we assume that these coefficients are returned by the function "GetQuadricCoeffs", given the noisy incomplete matrix $\widetilde{\boldsymbol{D}}$, the observation mask $\boldsymbol{W}$ and the dimensionality d. The global minimizer of (2.89) can be found analytically by calculating the roots of its derivative (a cubic). The process is then repeated over all coordinates k, and points i, until convergence. The resulting algorithm is remarkably simple, yet empirically converges fast. It naturally lends itself to a distributed implementation. We spell it out in Algorithm 2.4.

When applied to a large dataset of random, noiseless and complete distance matrices, Algorithm 2.4 converges to the global minimum of (2.88) in more than 99% of the cases [162].

2.5.3 Semidefinite Programming

Recall the characterization of EDMs (2.74) in Theorem 2.2. It states that $\boldsymbol{D}$ is an EDM if and only if the corresponding geometrically centered Gram matrix $-\frac{1}{2}\boldsymbol{J}\boldsymbol{D}\boldsymbol{J}$ is positive-semidefinite. Thus, it establishes a one-to-one correspondence between the cone of EDMs, denoted by $\mathbb{EDM}^n$, and the intersection of the symmetric positive-semidefinite cone $\mathbb{S}_+^n$ with the geometrically centered cone $\mathbb{S}_c^n$. The latter is defined as the set of all symmetric matrices whose column sum vanishes,

$$\mathbb{S}_c^n = \Big\{ \boldsymbol{G} \in \mathbb{R}^{n\times n} \mid \boldsymbol{G} = \boldsymbol{G}^\top,\ \boldsymbol{G}\boldsymbol{1} = \boldsymbol{0} \Big\}. \tag{2.90}$$

---

[2] While the experimental setup of Glunt [78] is not detailed, it was mentioned that the EDMs were produced randomly.

Algorithm 2.4 Alternating Descent [162]

```
function AlternatingDescent($\tilde{\boldsymbol{D}}$, $\boldsymbol{W}$, $d$)
    $\boldsymbol{X} \in \mathbb{R}^{d\times n} \leftarrow \boldsymbol{X}_0 = \boldsymbol{0}$    ▷ Initialize the point set
    repeat
        for $i \in \{1, \cdots, n\}$ do    ▷ Points
            for $k \in \{1, \cdots, d\}$ do    ▷ Coordinates
                $\boldsymbol{\alpha}^{(i,k)} \leftarrow$ GetQuadricCoeffs($\boldsymbol{W}$, $\tilde{\boldsymbol{D}}$, $d$)
                $x_{i,k} \leftarrow \arg\min_x f(x; \boldsymbol{\alpha}^{(i,k)})$    ▷ Eq. (2.89)
            end for
        end for
    until Convergence or MaxIter
    return $\boldsymbol{X}$
end function
```

We can use this correspondence to cast EDM completion and approximation as semidefinite programs. While (2.74) describes an EDM of an n-point configuration in any dimension, we are often interested in situations where $d \ll n$. It is easy to adjust for this case by requiring that the rank of the centered Gram matrix be bounded. One can verify that

$$\left.\begin{array}{r}\boldsymbol{D} = \mathrm{EDM}(\boldsymbol{X})\\ \mathrm{affdim}(\boldsymbol{X}) \leqslant d\end{array}\right\} \iff \left\{\begin{array}{l}-\frac{1}{2}\boldsymbol{J}\boldsymbol{D}\boldsymbol{J} \succeq 0\\ \mathrm{rank}(\boldsymbol{J}\boldsymbol{D}\boldsymbol{J}) \leqslant d,\end{array}\right. \tag{2.91}$$

when $n \geqslant d$. That is, EDMs with a particular embedding dimension d are completely characterized by the rank and definiteness of $\boldsymbol{J}\boldsymbol{D}\boldsymbol{J}$.

Now we can write the following rank-constrained semidefinite program for solving Problem 2.1,

$$\begin{array}{ll}\underset{\boldsymbol{G}}{\text{minimize}} & \| \boldsymbol{W} \circ \left(\tilde{\boldsymbol{D}} - \mathcal{K}(\boldsymbol{G})\right)\|_F^2\\ \text{subject to} & \mathrm{rank}(\boldsymbol{G}) \leqslant d\\ & \boldsymbol{G} \in \mathbb{S}_+^n \cap \mathbb{S}_c^n.\end{array} \tag{2.92}$$

The second constraint is just a shorthand for writing $\boldsymbol{G} \succeq 0,\ \boldsymbol{G}\boldsymbol{1} = \boldsymbol{0}$. We note that this is equivalent to MDS with the s-stress cost function, thanks to the rank characterization (2.91).

Unfortunately, the rank property makes the feasible set in (2.92) non-convex, and solving it exactly becomes difficult. This makes sense, as we know that s-stress is not convex. Nevertheless, we may *relax* the hard problem, by simply omitting the rank constraint, and hope to obtain a solution with the correct dimensionality,

$$\begin{array}{ll}\underset{\boldsymbol{G}}{\text{minimize}} & \| \boldsymbol{W} \circ \left(\tilde{\boldsymbol{D}} - \mathcal{K}(\boldsymbol{G})\right)\|_F^2\\ \text{subject to} & \boldsymbol{G} \in \mathbb{S}_+^n \cap \mathbb{S}_c^n.\end{array} \tag{2.93}$$

We call (2.93) a semidefinite relaxation (SDR) of the rank-constrained program (2.92).

The constraint $\boldsymbol{G} \in \mathbb{S}_c^n$, or equivalently, $\boldsymbol{G}\boldsymbol{1} = \boldsymbol{0}$, means that there are no strictly positive definite solutions ($\boldsymbol{G}$ has a nullspace, so at least one eigenvalue must be zero). In other words, there exist no strictly feasible points [113]. This may pose a numerical problem, especially for various interior point methods. The idea is then to reduce the size of the Gram matrix through an invertible transformation, somehow removing the part of it responsible for the nullspace. In what follows, we describe how to construct this smaller Gram matrix.

A different, equivalent way to phrase the multiplicative characterization (2.74) is the following statement: a symmetric hollow matrix $\boldsymbol{D}$ is an EDM if and only if it is negative semidefinite on $\{\boldsymbol{1}\}^\perp$ (on all vectors $\boldsymbol{t}$ such that $\boldsymbol{t}^\top \boldsymbol{1} = 0$). Let us construct an orthonormal basis for this orthogonal complement—a subspace of dimension $(n-1)$—and arrange it in the columns of matrix $\boldsymbol{V} \in \mathbb{R}^{n \times (n-1)}$. We demand

$$\begin{aligned} \boldsymbol{V}^\top \boldsymbol{1} &= \boldsymbol{0} \\ \boldsymbol{V}^\top \boldsymbol{V} &= \boldsymbol{I}. \end{aligned} \tag{2.94}$$

There are many possible choices for $\boldsymbol{V}$, but all of them obey that $\boldsymbol{V}\boldsymbol{V}^\top = \boldsymbol{I} - \frac{1}{n}\boldsymbol{1}\boldsymbol{1}^\top = \boldsymbol{J}$. The following choice is given in [4],

$$\boldsymbol{V} = \begin{bmatrix} p & p & \cdots & p \\ 1+q & q & \cdots & q \\ q & 1+q & \cdots & q \\ \vdots & \cdots & \ddots & \vdots \\ q & q & \cdots & 1+q \end{bmatrix}, \tag{2.95}$$

where $p = -1/(n+\sqrt{n})$ and $q = -1/\sqrt{n}$.

With the help of the matrix $\boldsymbol{V}$, we can now construct the sought Gramian with reduced dimensions. For an EDM $\boldsymbol{D} \in \mathbb{R}^{n \times n}$,

$$\mathcal{G}(\boldsymbol{D}) \overset{\text{def}}{=} -\frac{1}{2}\boldsymbol{V}^\top \boldsymbol{D}\boldsymbol{V} \tag{2.96}$$

is an $(n-1) \times (n-1)$ PSD matrix. This can be verified by substituting (2.96) in (2.67). Additionally, we have that

$$\mathcal{K}(\boldsymbol{V}\mathcal{G}(\boldsymbol{D})\boldsymbol{V}^\top) = \boldsymbol{D}. \tag{2.97}$$

Indeed, $\boldsymbol{H} \mapsto \mathcal{K}(\boldsymbol{V}\boldsymbol{H}\boldsymbol{V}^\top)$ is an invertible mapping from $\boldsymbol{S}_+^{n-1}$ to $\mathbb{EDM}^n$ whose inverse is exactly $\boldsymbol{G}$. Using these notations we can write down an equivalent optimization program that is numerically more stable than (2.93) [4]:

$$\begin{aligned} &\underset{\boldsymbol{H}}{\text{minimize}} && \|\boldsymbol{W} \circ \left(\tilde{\boldsymbol{D}} - \mathcal{K}(\boldsymbol{V}\boldsymbol{H}\boldsymbol{V}^\top)\right)\|_F^2 \\ &\text{subject to} && \boldsymbol{H} \in \mathbb{S}_+^{n-1}. \end{aligned} \tag{2.98}$$

On the one hand, with the above transformation the constraint $\boldsymbol{G}\boldsymbol{1} = \boldsymbol{0}$ became implicit in the objective, as $\boldsymbol{V}\boldsymbol{H}\boldsymbol{V}^\top\boldsymbol{1} \equiv \boldsymbol{0}$ by (2.94); on the other hand, the feasible set is now the full semidefinite cone $\mathbb{S}_+^{n-1}$.

Still, as Krislock & Wolkowicz mention [113], by omitting the rank constraint we allow the points to move about in a larger space, so we may end up with a higher-dimensional solution even if there is a completion in dimension d.

There exist various heuristics for promoting lower rank. One such heuristic involves the trace norm—the convex envelope of rank. The trace or nuclear norm is studied extensively by the compressed sensing community. In contrast to the common wisdom in compressed sensing, the trick here is to maximize the trace norm, not to minimize it. The mechanics are as follows: maximizing the sum of squared distances between the points will stretch the configuration as

Algorithm 2.5 Semidefinite Relaxation (Matlab/CVX)

```
function [EDM, X] = sdr_complete_edm(D, W, lambda)

n = size(D, 1);
x = -1/(n + sqrt(n));
y = -1/sqrt(n);
V = [y*ones(1, n-1); x*ones(n-1) + eye(n-1)];
e = ones(n, 1);

cvx_begin sdp
    variable G(n-1, n-1) symmetric;
    B = V*G*V';
    E = diag(B)*e' + e*diag(B)' - 2*B;
    maximize trace(G) ...
           - lambda * norm(W .* (E - D), 'fro');
    subject to
        G >= 0;
cvx_end

[U, S, V] = svd(B);
EDM = diag(B)*e' + e*diag(B)' - 2*B;
X = sqrt(S)*V';
```

much as possible, subject to available constraints. But stretching favors smaller affine dimensions (imagine pulling out a roll of paper, or stretching a bent string). Maximizing the sum of squared distances can be rewritten as maximizing the sum of norms in a centered point configuration—but that is exactly the trace of the Gram matrix $\boldsymbol{G} = -\frac{1}{2}\boldsymbol{J}\boldsymbol{D}\boldsymbol{J}$ [220]. This idea has been successfully put to work by Weinberger and Saul [220] in manifold learning, and by Biswas *et al.* in SNL [17].

Noting that $\operatorname{trace}(\boldsymbol{H}) = \operatorname{trace}(\boldsymbol{G})$ because $\operatorname{trace}(\boldsymbol{J}\boldsymbol{D}\boldsymbol{J}) = \operatorname{trace}(\boldsymbol{V}^\top \boldsymbol{D}\boldsymbol{V})$, we write the following SDR,

$$\begin{array}{ll} \underset{\boldsymbol{H}}{\text{maximize}} & \operatorname{trace}(\boldsymbol{H}) - \lambda \| \boldsymbol{W} \circ \big(\widetilde{\boldsymbol{D}} - \mathcal{K}(\boldsymbol{V}\boldsymbol{H}\boldsymbol{V}^\top)\big) \|_F \\ \text{subject to} & \boldsymbol{H} \in \mathbb{S}_+^{n-1}. \end{array} \tag{2.99}$$

Here we opted to include the data fidelity term in the Lagrangian form, as proposed by Biswas [17], but it could also be moved to constraints. Finally, in all of the above relaxations, it is straightforward to include upper and lower bounds on the distances. Because the bounds are linear constraints, the resulting programs remain convex; this is particularly useful in the molecular conformation problem. A Matlab/CVX [85, 84] implementation of the SDR (2.99) is given in Algorithm 2.5.

2.5.4 Performance Comparison of Algorithms

We compare the described algorithms on the task of EDM completion. The entries to delete are chosen uniformly at random. In Figure 2.15, we assume that the observed entries are known exactly, and we plot the success rate (percentage of accurate EDM reconstructions) against the number of deletions. Accurate reconstruction is defined in terms of the relative error. Let $\boldsymbol{D}$ be the true, and $\widehat{\boldsymbol{D}}$ the estimated EDM. The relative error is then $\|\widehat{\boldsymbol{D}} - \boldsymbol{D}\|_F / \|\boldsymbol{D}\|_F$, and we declare success if this error is below 1%.

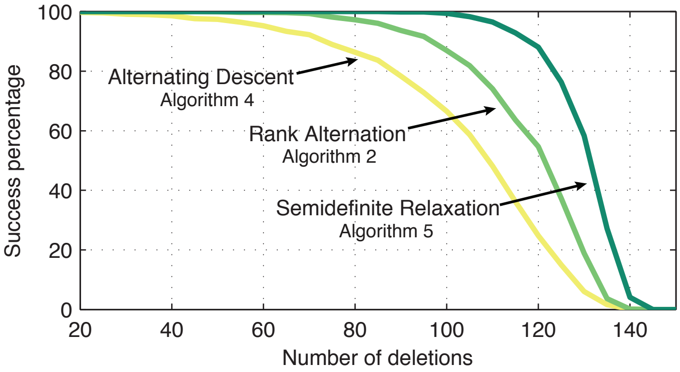


Figure 2.15: Comparison of different algorithms applied to completing an EDM with random deletions. For every number of deletions, we generated 2000 realizations of 20 points uniformly at random in a unit square. Distances to delete were chosen uniformly at random among the resulting $\binom{20}{2} = 190$ pairs; 20 deletions correspond to $\approx$ 10% of the number of distance pairs and to 5% of the number of matrix entries; 150 deletions correspond to $\approx$ 80% of the distance pairs and to $\approx$ 38% of the number of matrix entries. Success was declared if the Frobenius norm of the error between the estimated matrix and the true EDM was less than 1% of the Frobenius norm of the true EDM.

To generate Figure 2.16 we varied the amount of random, uniformly distributed jitter added to the distances, and for each jitter level we plotted the relative error. The exact values of intermediate curves are less important than the curves for the smallest and the largest jitter, and the overall shape of the ensemble.

A number of observations can be made about the performance of algorithms. Notably, OptSpace (Algorithm 2.3) does not perform well for randomly deleted entries when $n = 20$; it was designed for larger matrices. For this matrix size, the mean relative reconstruction error achieved by OptSpace is the worst of all algorithms (Figure 2.16). In fact, the relative error in the noiseless case was rarely below the success threshold (set to 1%) so we omitted the corresponding near-zero curve from Figure 2.15. To fully exploit OptSpace, n should be much larger, in the thousands or tens of thousands.

SDR (Algorithm 2.5) performs well in all scenarios. For both the random deletions and the MDU, it has the highest success rate, and it behaves well with respect to noise. Alternating coordinate descent (Algorithm 2.4) performs slightly better in noise for small number of deletions and large number of calibration events, but Figure 2.15 indicates that for certain realizations of the point set it gives large errors. If the worst-case performance is critical, SDR is a better choice. We note that in the experiments involving the SDR, we have set the multiplier λ in (2.99) to the square root of the number of missing entries. This choice was empirically found to perform well; performance could be improved by a more careful choice (in particular, by making it a function of the noise level).

The main drawback of SDR is speed; it is the slowest among the tested algorithms. To solve the semidefinite program we used CVX [85, 84], a Matlab interface to various interior point methods. For larger matrices (*e.g.*, $n = 1000$), CVX runs out of memory on a desktop computer, and essentially never finishes. Matlab implementations of alternating coordinate descent, rank alternation (Algorithm 2.2), and OptSpace are all much faster.

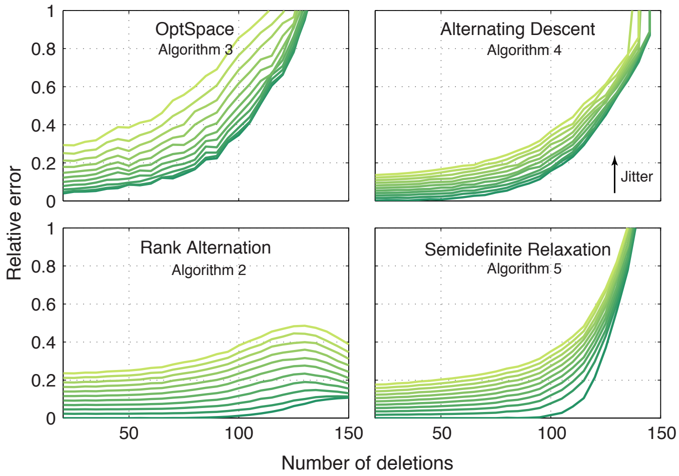


Figure 2.16: Comparison of different algorithms applied to completing an EDM with random deletions and noisy distances. For every number of deletions, we generated 1000 realizations of 20 points uniformly at random in a unit square. In addition to the number of deletions, we varied the amount of jitter added to the distances. Jitter was drawn from a centered uniform distribution, with the level increasing in the direction of the arrow, from $\mathcal{U}[0,0]$ (no jitter) for the darkest curve at the bottom, to $\mathcal{U}[-0.15,0.15]$ for the lightest curve at the top, in 11 increments. For every jitter level, we plotted the mean relative error $\|\widehat{\boldsymbol{D}}-\boldsymbol{D}\|_F/\|\boldsymbol{D}\|_F$ for all algorithms.

To summarize, for smaller matrices the SDR seems to be the best overall choice. For large matrices the SDR becomes too slow and one should turn to alternating coordinate descent, rank alternation or OptSpace. Rank alternation is the simplest algorithm, but alternating coordinate descent performs better. For very large matrices (n on the order of thousands or tens of thousands), OptSpace becomes the most attractive solution. We note that we deliberately refrained from making detailed running time comparisons, due to the diverse implementations of the algorithms.

2.6 Conclusion

Two key concepts introduced in this chapter are the image source model and Euclidean distance geometry. We will see in the following chapters how, when appropriately combined, they give rise to conceptually simple and robust algorithms for solving various geometric inverse problems in room acoustics (and beyond). This chapter has laid out the material and properties we will need to do so effectively, such as the results on rank and essential uniqueness of EDMs, and algorithms to reconstruct point configurations from noisy, incomplete sets of distances. The trick will always be the same: Use the Euclidean distance geometry to *lift* the distance information extracted from 1D room impulse responses to 3D points that are the image sources.

Chapter 3

Can One Hear the Shape of a Room?*

> What do you see when you turn out the light?
> I can't tell you, but I know it's mine.
>
> *The Beatles*

3.1 Introduction

In a famous paper [99], Mark Kac asks the catchy question "Can one hear the shape of a drum?". More concretely, he asks whether two membranes of different shapes necessarily resonate at different frequencies.[1] This problem is related to a question in astrophysics [80], and the answer turns out to be negative: Using tools from group representation theory, Gordon, Webb and Wolpert [81, 82] presented several elegantly constructed counterexamples, including the two polygonal drum shapes shown in Figure 3.1. Although geometrically distinct, the two drums have the same resonant frequencies.

We ask a similar question about rooms. Assume you are blindfolded inside a room; you snap your fingers and listen to echoes. Can you hear the shape of the room? Intuitively, and for simple room shapes, we know that this is possible. A shoebox room, for example, has well-defined modes, from which we can derive its size. But the question is challenging in more general cases, even if we presume that the room impulse response (RIR) contains an arbitrarily long sequence of echoes (with an ideal, noiseless measurement) that should ultimately specify the room geometry.

---

[1] Resonant frequencies correspond to the square root of the eigenvalues of the Laplacian on a 2D domain.

[*] This chapter describes results from joint works with Reza Parhizkar, Andreas Walther, Yue M. Lu, and Martin Vetterli [54, 56].

It might appear that Kac's problem and the question we pose are equivalent. This is not the case, for the sound of a drum depends on more than its set of resonant frequencies (eigenvalues)—it also depends on its resonant modes (eigenvectors). In the paper "Drums that sound the same" [34], Chapman explains how to construct drums of different shapes with matching resonant frequencies. Still, these drums would hardly *sound* the same if hit with a drumstick. They share the resonant frequencies, but the impulse responses are different, and we want to hear rooms from impulse responses. Even a single drum struck at different points sounds differently. This is shown clearly in Figure 1.

Another difference is that we are not interested in the uniqueness for all instances. We only want to show that, generically, most of the rooms that satisfy our modeling assumptions can be *heard*; if rooms or microphone locations are chosen at random, we want the uniqueness to be almost sure. Differently from Kac, we are more interested in algorithms that reconstruct rooms from echoes.

Certain animals can indeed "hear" their environment. Bats, dolphins and some birds probe the environment by emitting sounds and then use echoes to navigate. Remarkably, there are people that can do the same, or better. Daniel Kish produces clicks with his mouth, and uses echoes to learn the shape, distance and density of objects around him [179]. The main cues for human echolocators are the early reflections; our computer algorithms also use early reflections to calculate shapes of rooms.

Many applications benefit from knowing the room geometry. Indoor sound source localization is usually considered difficult, because the reflections are difficult to predict and they masquerade as sources. Yet, somewhat surprisingly, in rooms one can do localization more accurately than in free-field if the room geometry is known [8, 175, 177, 61]. In teleconferencing, auralization and virtual reality, one often needs to compensate the room influence or create an illusion of a specific room. The success of these tasks largely depends on the accurate modeling of the early reflections [128], which in turn requires the knowledge of where the walls are.

All but a few works (which appeared after our paper written in 2010 and presented at ICASSP 2011 [54]) employ array processing methods to reconstruct the shape of a room. The few works that attempt to do it from a single RIR restrict themselves to simple scenarios. Marković *et al.* [138] discuss the estimation of the geometry of rectangular and L-shaped rooms from a single impulse response. They use a genetic algorithm to address the global optimization problem with many local minima; this approach does not allow them to analyze the problem theoretically or to provide guarantees about the uniqueness of the solution. Furthermore, it is not possible to guarantee that the estimated solution is the correct one, even in the noiseless case.

Moore, Brookes, and Naylor [152] propose to estimate the geometry of a 2D rectangular room from one impulse response using the regularities in the times of arrivals of echoes. For a simple rectangular room, these have simple expressions that are functions of the lengths of the room sides, which is leveraged by their least-squares algorithm.

In this chapter, we first discuss room estimation from a single measured RIR, that is, by only using a single static microphone. This part of the discussion is primarily theoretical; our goal is to show that a single room impulse response is sufficient to reconstruct convex polyhedral rooms in theory, although it presents certain challenges in practice. We first assume that we receive echoes from all the first-order and second-order image sources, and then show how this assumption can be relaxed.

Next, we show how to reconstruct a convex polyhedral room from a few impulse responses. When using more than one microphone, many of the assignment challenges from the single-microphone case disappear, mostly because we can restrict ourselves to first-order echoes. Our method relies on learning from which wall a particular echo originates. There are two challenges with this approach: First, it is difficult to extract echoes from RIRs; and second, the microphones receive echoes from walls in different orders. The main novelty in our approach is an algorithm that selects the “correct” combinations of echoes, specifically those that actually correspond to walls. The need for assigning echoes to walls arises from the omnidirectionality of the source and the receivers. The assignment problem is combinatorial (though of a small size) in the general case of noisy distances. For small noise, however, we show that it can be solved in polynomial time.

There have been several attempts to estimate the room geometry from RIRs. In [7], the problem is formulated in 2D, and the authors take advantage of multiple source locations to estimate the geometry. In [176] the authors address the problem by ℓ_1-regularized template matching with a measured dictionary of impulse responses. Their approach requires measuring a very large matrix of impulse responses for a fixed source-receiver geometry. The authors in [200] propose a 3D room reconstruction method by assuming that the array is small enough so that there is no need to assign echoes to walls. They use sparse RIRs obtained by directing the loudspeaker to many orientations and processing the obtained responses. In contrast, our method works with arbitrary measurement geometries. Furthermore, we prove that the first-order echoes provide a unique description of the room for almost all setups. A subspace-based formulation allows us to use the minimal number of microphones (4 microphones in 3D). It is impossible to further reduce the number of microphones, unless we consider higher-order echoes, as attempted in [54]. However, the arrival times of higher-order echoes are often challenging to obtain and delicate to use, both for theoretical and for practical reasons. Therefore, in the proposed method, we choose to use more than one microphone, thus avoiding the need for higher-order echoes.

In addition to theoretical analysis, we validate the results experimentally by “hearing” rooms on EPFL campus. Moreover, by running it in an alcove of the Lausanne cathedral, we show that the algorithm still gives a useful output even when the room thoroughly violates the assumptions of it being a convex polyhedron.

3.2 Can One Hear the Shape of a Drum?

Kac’s 1966 question is a question about well-posedness of the “shape from resonant frequencies” inverse problem. More concretely, it is a question about uniqueness of the solution: Are there geometrically distinct drums with the same infinity of resonant frequencies? If there were none, it would imply drum *hearability* provided that we have a good algorithm or a very powerful computer that could solve the inverse spectral problem.

Let us state Kac’s question more precisely. It is known that the oscillations of a membrane of shape Ω are governed by the wave equation,

$$\begin{aligned} &\Delta u - \frac{1}{c^2}\frac{\partial^2 u}{\partial t^2} = 0 \text{ inside } \Omega \\ &u = 0 \text{ on } \partial\Omega. \end{aligned} \tag{3.1}$$

We can try to solve (3.1) by separating the spatial and the temporal variables,

$$u(\boldsymbol{x}, t) = \nu(\boldsymbol{x})\psi(t). \tag{3.2}$$

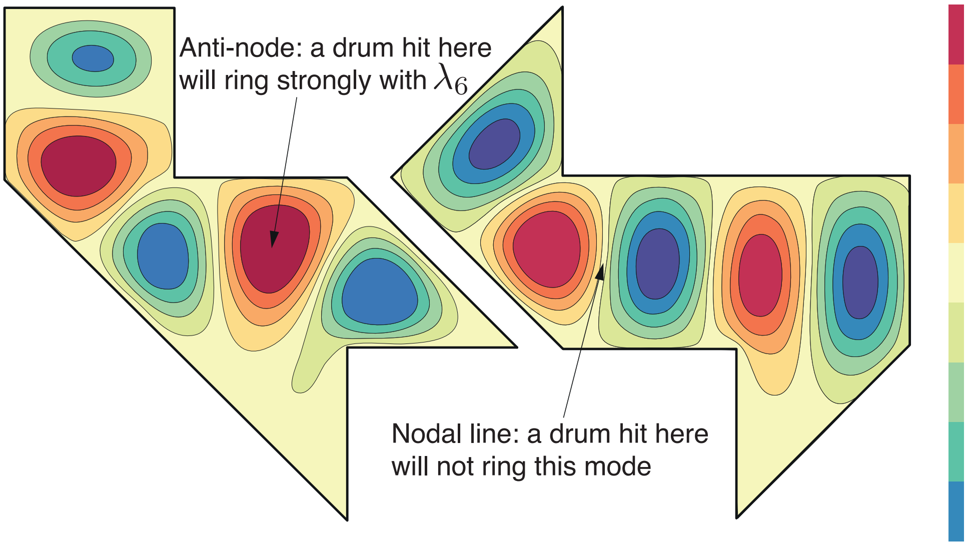

Anti-node: a drum hit here will ring strongly with λ_6
Nodal line: a drum hit here will not ring this mode

Figure 3.1: Figure shows two isospectral drums [66]. Although geometrically distinct, these drums have the same resonant frequencies. The standing wave corresponding to the eigenvalue λ_6 is shown for both drums. It is clear that this mode will be excited with different amplitudes, depending on where we hit the drum. Extremes are nodes and anti-nodes.

Plugging this into (3.1), we can group together the spatial and the temporal parts,

$$\frac{\Delta\nu(\boldsymbol{x})}{\nu(\boldsymbol{x})} = \frac{1}{c^2}\frac{\partial^2\psi(t)}{\partial t^2}\Big/\psi(t). \tag{3.3}$$

The left-hand side does not change with t, so the right-hand side must be a constant. Similarly, the left-hand side must be a constant too, so the two sides must equal the same constant. Calling this constant λ, we get for the spatial part an eigenvalue problem that

$$\begin{aligned} &\Delta\nu(\boldsymbol{x}) = \lambda\nu(\boldsymbol{x}) \text{ inside } \Omega \\ &\nu(\boldsymbol{x}) = 0 \text{ on } \partial\Omega, \end{aligned} \tag{3.4}$$

which has a solution only for specific values of λ—the Dirichlet eigenvalues of the Laplacian on Ω. In fact, for compact Ω, it is known that there is a discrete set of eigenvalues for which (3.4) admits a solution, and this set is called the *spectrum* of the Laplacian Δ on Ω.

Kac's question can then be asked precisely as follows. Let Ω_1 and Ω_2 be two (open) plane regions bounded by curves $\partial\Omega_1$ and $\partial\Omega_2$. Consider the two eigenvalue problems

$$\Delta\nu(\boldsymbol{x}) = \lambda\nu(\boldsymbol{x}) \text{ for } \boldsymbol{x} \in \Omega_1 \qquad \Delta\mu(\boldsymbol{x}) = \kappa\mu(\boldsymbol{x}) \text{ for } \boldsymbol{x} \in \Omega_2 \tag{3.5}$$

$$\nu(\boldsymbol{x}) = 0 \text{ for } \boldsymbol{x} \in \partial\Omega_1 \qquad \mu(\boldsymbol{x}) = 0 \text{ for } \boldsymbol{x} \in \partial\Omega_2 \tag{3.6}$$

and assume that for each n, the sorted eigenvalue λ_n for Ω_1 equals the sorted eigenvalue κ_n for Ω_2. Does it imply that Ω_1 and Ω_2 are congruent in the sense of Euclidean geometry?

3.2.1 Historical Notes and Kac's Paper

In his paper, Kac attributes the catchy phrasing of the question to Lipman Bers, and the question itself to Salomon Bochner. We mentioned in the introduction to this chapter that this problem is related to a problem in astrophysics; in fact the astrophysics reference predates the Kac/Bers/Bochner question about drums by many decades.

Sir Arthur Schuster (1851–1934) is credited with being one of the founding fathers of astrophysics. He coined the term *spectroscopy*, and he is believed to be the first one to put forward the idea of solving inverse problems of atomic and molecular structure using spectroscopy. At the time, molecular spectra were used only to infer the chemical composition of, for example, the Sun. But he coined the term spectroscopy to refer exactly to the problem of learning about the structure of atoms and molecules from spectra they can generate. In his 1882 report to the British Association, entitled "The Genesis of Spectra", Schuster wrote

> "We know a great deal more about the forces which produce the vibrations of sound than about those which produce the vibrations of light. To find out the different tunes sent out by a vibrating system is a problem which may or may not be solvable in certain special cases, but it would baffle the most skillful mathematician to solve the inverse problem and to find out the shape of a bell by means of the sounds which it is capable of sending out. And this is the problem which ultimately spectroscopy hopes to solve in the case of light. In the meantime we must welcome with delight even the smallest step in the desired direction."

So Schuster posed a question about bells very similar to the more famous question about drums already in 1882!

There is an interesting anecdote related to Schuster's question. The famous French philosopher Auguste Comte (1798–1857) wrote in his magnum opus "Cours de la Philosophie Positive" in 1835 [39, 40] that

> "It is easy to describe clearly the character of astronomical science, from its being thoroughly separated, in our time, from all theological and metaphysical influence. Looking at the simple facts of the case, it is evident that though three of our senses take cognizance of distant objects, only one of the three perceives the stars. The blind could know nothing of them; and we who see, after all our preparation, know nothing of stars hidden by distance, except by induction. Of all objects, the planets are those which appear to us under the least varied aspect. We see how we may determine their forms, their distances, their bulk, and their motions, but we can never know anything of their chemical or mineralogical structure; and, much less, that of organized beings living on their surface. We may obtain positive knowledge of their geometrical and mechanical phenomena; but all physical, chemical, physiological, and social researches, for which our powers fit us on our own earth, are out of the question in regard to the planets. Whatever knowledge is obtainable by means of the sense of Sight, we may hope to attain with regard to the stars, whether we at present see the method or not; and whatever knowledge requires the aid of other senses, we must at once exclude from our expectations, in spite of any appearances to the contrary."[2]

---

[2] "...il est indispensable, en sortant des définitions vagues qu'on en donne encore habituellement, de commencer par circonscrire avec exactitude le véritable champ des connaissances positives que nous pouvons acquérir à l'égard des astres. ... Toute recherche qui n'est point finalement réductible à de simples observations visuelles

Only two years after Comte's death in 1857, Gustav Kirchoff and Robert Bunsen discovered (in Bunsen's own words *unexpectedly*) the cause of the dark lines in the solar spectra, and set out to determine the chemical composition of the Sun.

Curiously, there is another bad prediction related to Kac's question, made by a scholar of equal stature as Comte. While discussing what we *can* hear from the spectrum, Kac mentions that one can *hear* the area of a drum (or in general, the volume of a region). This is the subject of a conjecture by Hendrik Lorentz which he posed while lecturing at the end of 1910 in Göttingen. It was known at the time that the result holds in many particular cases, and Lorentz correctly believed that it holds in general. David Hilbert predicted that it would not be proved during his lifetime, but only two years later, his student Hermann Weyl proved it using the tools of integral equations which were developed precisely by Hilbert several years earlier.

3.2.2 Counterexamples: Isospectral Drums

Kac's question was answered in 1992 by Gordon, Webb and Wollpert, in an announcement in the Bulletin of the American mathematical society appropriately titled "One Cannot Hear the Shape of a Drum" [82] and the accompanying full paper [81]. They exhibited *isospectral* drums with the same set of resonant frequencies, or spectrum. We note that the actual resonant frequency is equal to the square root of the eigenvalue λ in (3.4).

The reader might be interested by the construction of a counterexample. Here, we explain one given by Gordon & Webb in [80]. The beauty of these counterexamples is that they can be understood by elementary means; but how to systematically arrive at such constructions is more involved and requires tools from group representation theory [81].

To understand the counterexample we will use two properties of the solutions to the eigenvalue problem (3.4) (which is actually a Helmholtz equation):

(i) *Linearity*: Linear combination of solutions is again a solution,

(ii) *Reflection principle*: If we have a solution on a domain bounded by a straight line segment with the clamped (Dirichlet) boundary condition, we can extend the domain and the solution by mirroring it over the line segment and changing the sign. This procedure ensures that the solution continues smoothly into the mirrored domain.

Now consider the two drumheads in Figure 3.2 (of the same shape as in Figure 3.1). The drums are segmented and annotated as in [80]. Let the vibrations of D1 be described by a function φ supported on the drum. The function φ satisfies both the equation and the boundary condition for a given λ. Let $A, B, \ldots, G$ denote the restrictions of φ to corresponding triangular segments as indicated in the figure.

There happens to be a way to "transplant" the waveform from D1 to D2, so that the resulting waveform on D2 still satisfies the equation and the boundary condition. This transplantation is effected by placing linear combinations of $A, B, \ldots, G$ on D2, as indicated in Figure 3.2, while observing the edge colors to ensure proper orientations.

---

nous est donc nécessairement interdite au sujet des astres, qui sont ainsi de tous les êtres naturels ceux que nous pouvons connaître sous les rapports les moins variés. Nous concevons la possibilité de déterminer leurs formes, leurs distances, leurs grandeurs et leurs mouvemens; tandis que nous ne saurions jamais étudier par aucun moyen leur composition chimique, ou leur structure minéralogique, et, à plus forte raison, la nature des corps organisés qui vivent à leur surface, etc. En un mot, pour employer immédiatement les expressions scientifiques les plus précises, nos connaissances positives par rapport aux astres sont nécessairement limitées à leurs seuls phénomènes géométriques et mécaniques, sans pouvoir nullement embrasser les autres recherches physiques, chimiques, physiologiques, et même sociales, que comportent les êtres accessibles à tous nos divers moyens d'observation."

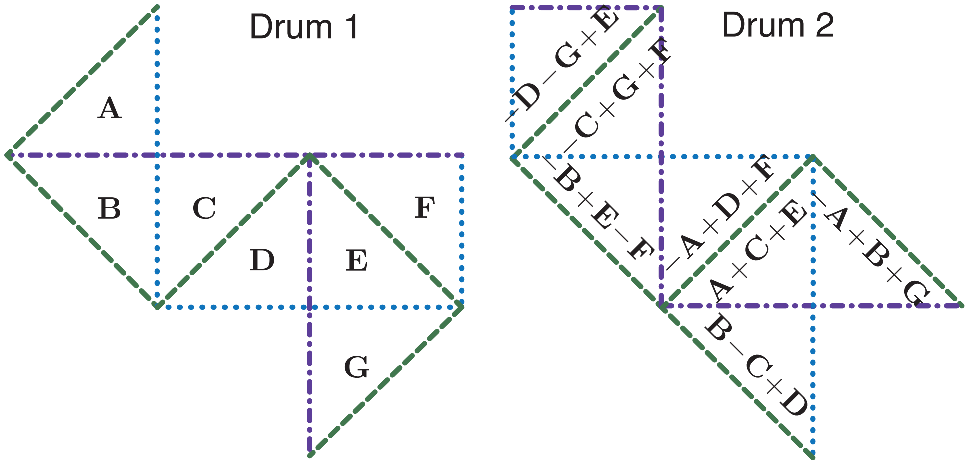


Figure 3.2: Isospectral drums with annotations as in [80].

We can check that the transplanted waveform indeed satisfies both the equation and the boundary condition. Consider for example triangles $A + C + E$ and $-A + D + F$ on D2. We require that the corresponding waveforms combine smoothly over the green edge. Triangles C and D share the green edge on D1. The same holds for triangles E and F. This means that they combine smoothly on D1 so $C + E$ and $D + F$ will combine smoothly on D2 as well. Now observe that the green edge of A on D1 is the boundary edge, so A vanishes along the green edge. By reflection principle we can continue A smoothly over the green edge by mirroring it and multiplying by -1. Finally, this implies that $A + C + E$ and $-A + D + F$ will stitch smoothly. In order to check that the boundary conditions are satisfied, consider for example the triangle $-A + B + G$ and its purple boundary edge. Triangles A and B share the purple edge in D1, so they necessarily have the same value on the purple edge. Thus $-A + B$ is zero over the edge. In triangle G, the purple edge is the boundary edge, so G is zero on that edge, and $-A+B+G$ must be zero on the boundary edge. This check can be effected for all triangles in D2. We finally note that the equation is satisfied simply because linear combinations of solutions are again solutions.

We showed that the equation (3.4) holds with the same λ for both drums. Therefore every resonant mode of D1 is also a resonant mode of D2. As we can also do a reverse transplantation procedure, every resonant mode of D2 is a resonant mode of D1, thus the two sets coincide, and the drums are isospectral.

3.3 Unlabeled Distances

Gordon, Webb and Wollpert have shown that a drum cannot be *heard* by listening to its resonances. When we ask "Can One Hear the Shape of a Room?", we want to know whether it is possible to reconstruct the room's shape from impulse responses. It turns out that to do it, we need to know how to create correspondences between the distances (peaks in the impulse responses) and the image sources (points in space). In other words, we must label the distances.

There are a number of applications in which we can measure the distances between the points, but we do not know the correct labeling. That is, we know all the entries of an EDM, but we

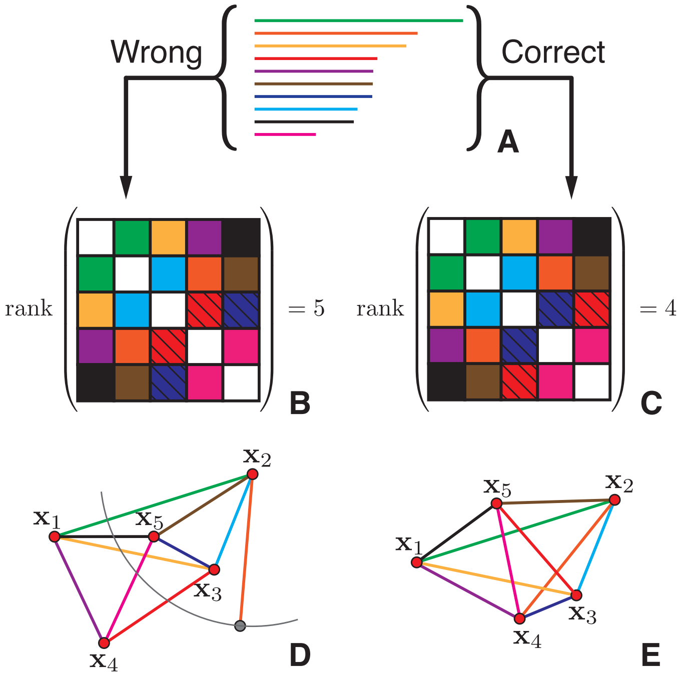


Figure 3.3: Illustration of the uniqueness of EDMs for unlabeled distances. A set of unlabeled distance (A) is distributed in two different ways in a tentative EDM with embedding dimension 2 (B and C). The correct assignment yields the matrix with the expected rank (C), and the point set is easily realized in the plane (E). On the contrary, swapping just two distances (hatched squares in (B) and (C)) makes it impossible to realize the point set in the plane (D). Triangles that do not coincide with the swapped edges can still be placed, but in the end we are left with a hanging orange stick that cannot attach itself to any of the five nodes.

do not know how to arrange them in the matrix. As illustrated in Figure 3.3, we can imagine having a set of sticks of various lengths. The task is to work out the correct way to connect the ends of the sticks so that no stick is left hanging open-ended.

It is interesting to note that in many cases, distance labeling is not essential: For most point configurations, there is no other set of points that can generate the corresponding set of distances, up to a rigid transformation. The labeling information is somehow implicit in the point set (Figure 3.3 illustrates exactly this point). If we assume a wrong labeling, things will not *click together*.

Localization from unlabeled distances is relevant in various calibration scenarios where we cannot tell apart the distance measurements belonging to different points in space. This can occur when we measure the times of arrivals of echoes, which correspond to distances between microphones and image sources [51, 56]. Somewhat surprisingly, the same problem of unlabeled distances appears in sparse phase retrieval.

No efficient algorithm currently exists for localization from unlabeled distances in the general case of noisy distances. We should mention, however, a recent polynomial-time algorithm (albeit of a high degree) by Gujarathi and *et al.* [87], that can reconstruct relatively large point sets from unordered, noiseless distance data.

At any rate, the number of assignments to test is sometimes sufficiently small so that an exhaustive search does not present a problem. We can then use EDMs to find the best labeling. The key to the unknown permutation problem is the following fact:

Theorem 3.1

*Draw $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n \in \mathbb{R}^d$ from some absolutely continuous probability distribution (*e.g.* uniformly at random) on $\Omega \subseteq \mathbb{R}^d$. Then with probability 1, the obtained point configuration is the unique (up to a rigid transformation) point configuration in Ω that generates the set of distances $\{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|, 1 \leqslant i < j \leqslant n\}$.*

This fact is a simple consequence of a result by Boutin and Kemper [23] who give a detailed characterization of point sets reconstructible from unlabeled distances. They show that the point sets that cannot be reconstructed from unordered distances (*i.e.*, those that generate distances that are also generated by other point sets) live on the zero set of a certain polynomial.

Figures 3.3B and 3.3C show two possible arrangements of a set of distances in a tentative EDM; the only difference is that the two hatched entries are swapped. But this simple swap is not harmless, as there is no way to attach the last stick in Figure 3.3D, while keeping the remaining triangles consistent. We could do it in a higher embedding dimension, but we insist on realizing it in the plane.

What Theorem 3.1 does not tell us is how to identify the correct labeling. But we know that for most sets of distances, only one (correct!) permutation can be realized in the given embedding dimension. Of course, if all the labelings are unknown and we have no good heuristics to trim the solution space, finding the correct labeling is difficult. Yet there are interesting situations where this search is feasible because we can augment the EDM point by point (see Section 3.8).

3.4 Modeling

Sound propagation in a room is described by a family of RIRs. An RIR models the channel between a fixed source and a fixed receiver. It contains the direct path and the reflections. Ideally, it is a train of pulses, each corresponding to an echo. For a fixed source and receiver positions it is given as

$$h(t) = \sum_i \alpha_i \delta(t - \tau_i), \tag{3.7}$$

and it changes with their positions.

A microphone *hears* the convolution of the emitted sound with the corresponding RIR, $y = x * h = \int x(s)h(\cdot - s)\, ds$. By measuring the impulse responses we access the propagation times τ_i, and these can be linked to the room geometry by the image source (IS) model discussed in Chapter 2. According to the IS model, we can replace reflections by virtual sources. Recall that virtual sources are simply mirror images of the true sources across the corresponding reflecting walls, and that the image $\widetilde{\boldsymbol{s}}_i$ of the source $\boldsymbol{s}$ with respect to the ith wall is computed as

$$\widetilde{\boldsymbol{s}}_i = \boldsymbol{s} + 2\,\langle \boldsymbol{p}_i - \boldsymbol{s}, \boldsymbol{n}_i \rangle\, \boldsymbol{n}_i, \tag{3.8}$$

where $\boldsymbol{n}_i$ is the unit normal corresponding to the ith wall, and $\boldsymbol{p}_i$ any point belonging to the ith wall. The time of arrival (TOA) of the echo from the ith wall is $t_i = \|\widetilde{\boldsymbol{s}}_i - \boldsymbol{r}\|/c$, where c is the speed of sound.

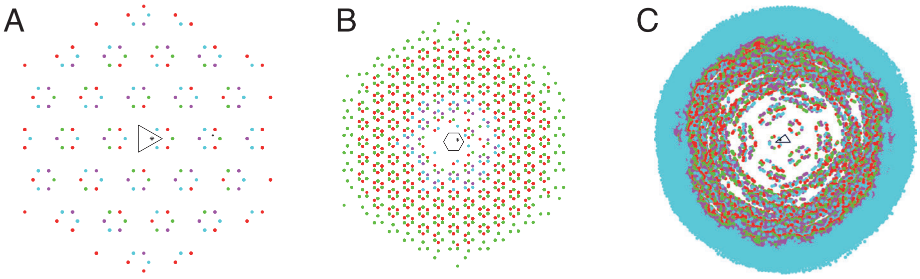


Figure 3.4: Image source patterns for different rooms. (A) equilateral triangle (regular pattern), (B) regular hexagon (regular pattern), (C) random triangle with 25 generations of virtual sources (irregular pattern). Since the patterns in (A) and (B) have *constant density*, the time density of the corresponding echoes scales as $\sim t$.

We should mention that measuring RIRs is an important topic in room acoustics, but we do not address it in this thesis. For details on how the experiments in this chapter were performed, please see the Supporting Information (SI) of [56].

3.5 Hearing the Shape of a Room with One Microphone

We first consider a setup that consists of one sound source and one colocated microphone, both omnidirectional. Assume that an omnidirectional pulse is emitted from somewhere inside the room, and that the room response is recorded at the same point. From the recorded RIR and the knowledge of the emitted pulse, we can extract the set of echo delays $\{\tau_i\}$.

For the purpose of this section, a room is a convex planar K-sided polygon represented by a $2 \times K$ vertex matrix $\boldsymbol{P} = [\boldsymbol{p}_1, \cdots, \boldsymbol{p}_K]$. For simplicity we operate in 2D, but it will be clear that the extension to 3D is straightforward. We assume that the vertices are specified in a counterclockwise direction and define the ith side of the room as the line segment joining $\boldsymbol{p}_i$ and $\boldsymbol{p}_{i+1}$. Without loss of generality we place the source at the origin and require that the room contains the origin. It is not possible to discriminate rotated and reflected versions of a room about the source; therefore, we think of these as being the same room. We can resolve this ambiguity by choosing (fixing) some degrees of freedom; *e.g.*, if the ith side is closest to the source, we set it to be vertical and choose that the closer of the adjacent sides follows in the CCW direction.

With the ith side of the room we associate an outward pointing unit normal $\boldsymbol{n}_i$, and we define the matrix of normals as $\boldsymbol{N} \overset{\text{def}}{=} [\boldsymbol{n}_1, \ldots, \boldsymbol{n}_K] \in \mathbb{R}^{2\times K}$. We denote by $\boldsymbol{q}_i$ the image source location with respect to the ith side. The set $\{\boldsymbol{q}_i\}_{1\leqslant i\leqslant K}$ then contains the first-order (or first-generation) image sources. Analogously, the image of the virtual source $\boldsymbol{q}_i$ with respect to the wall j is denoted $\boldsymbol{q}_{ij}$ so that $\{\boldsymbol{q}_{ij}\}_{1\leqslant i\neq j\leqslant K}$ is the set of second-order image sources.

By observing the impulse response we have access to $G_1 = \{\|\boldsymbol{q}_i\|\}$ and to $G_2 = \{\|\boldsymbol{q}_{ij}\|\}$—sets of times of arrivals of first- and second-order echoes. We assume everywhere that the speed of sound is $c = 1$ so that (numerically) distances equal times.

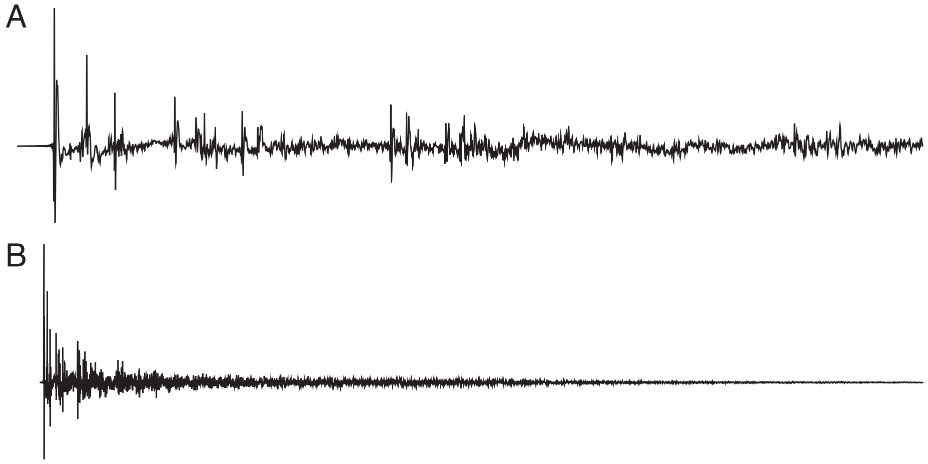


Figure 3.5: A room impulse response recorded in a classroom at EPFL (BC329). The early part (50 ms) with clearly distinguishable echoes is shown in (A). Early echoes together with the first part of the late reverberation (0.5 s) are shown in (B).

3.5.1 Characterization of RIRs from Image Source Patterns

Figure 3.4 shows the image source patterns for different rooms. From the generated image source patterns we can deduce certain known facts in room acoustics. One can observe a difference in the behavior of *regular* and *irregular* rooms.

For rectangular rooms the pattern of image sources corresponds to a union of four lattices in $\mathbb{R}^2$. A lattice $\Lambda_{\boldsymbol{M}} \in \mathbb{R}^2$ generated by matrix $\boldsymbol{M} \in \mathbb{R}^{2\times 2}$ is defined as $\Lambda_{\boldsymbol{M}} = \left\{\boldsymbol{x} | \boldsymbol{x} = \boldsymbol{M}\boldsymbol{n}, \boldsymbol{n} \in \mathbb{Z}^2\right\}$. From Figure 3.4A and Figure 3.4B we can observe that other regular polygons also generate regular image source patterns that correspond to unions of lattices. In contrast, Figure 3.4C shows that the image source pattern generated by a *random* triangle is not regular at all. An interesting effect is observed—as we move away from the original source, the density of the virtual sources increases. From these patterns we observe an interesting scaling law for regular rooms. The density of echoes in the RIR is growing with time as t^{d-1} where d is the dimensionality of the ambient space, *e.g.*, constant for 1D rooms, $\sim t$ in 2D rooms and $\sim t^2$ in 3D rooms. For a 2D room, this means that the number of image sources inside the annulus of constant width grows linearly with the radius of the annulus. If we consider the irregular triangle in Figure 3.4C it becomes clear that the same scaling statement does not hold.

We can distinguish between the sources corresponding to discrete early reflections, and the far sources that will generate the late reverberation—a known behavior from room acoustics. Early echoes and late reverberation are illustrated in Figure 3.5.

These examples suggest a very strong link between the room geometry—what we want to know—and the corresponding impulse response—what we hear. We show that under right conditions this link is an invertible mapping.

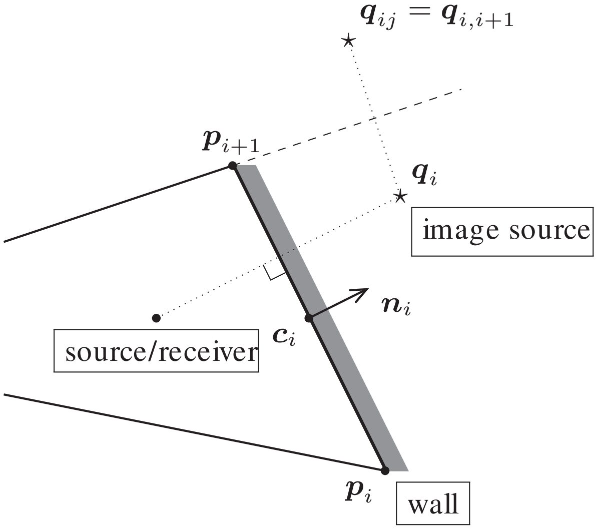


Figure 3.6: Setup with colocated source and receiver. Source is assumed to be at the origin. $\boldsymbol{p}_i$ and $\boldsymbol{p}_{i+1}$ are endpoints of ith wall, $\boldsymbol{n}_i$ is its unit, outward pointing normal, $\boldsymbol{c}_i$ is the center of the line segment $\boldsymbol{p}_i\boldsymbol{p}_{i+1}$, and $\boldsymbol{q}_i$ is the first generation image source. Its image with respect to the $(i+1)$st wall is $\boldsymbol{q}_{ij}$.

3.6 Room Geometry Estimation

In this section we derive the mapping between the room geometry and the RIR. We also discuss its uniqueness, and give an algorithm to retrieve the room geometry from the measured RIR.

3.6.1 The Shape of a Polygonal Room Using Matrix Analysis

It is not possible to reconstruct the room geometry using only the delays of the first-order echoes in G_1. To see this, consider a triangle with the corresponding set of first generation delays. Now choose one side and tilt it so that you change the shape of the triangle; this can only change one delay in G_1. But now we can translate this side keeping all the angles fixed until we match this changed delay with the old one, ending up with two rooms with the same G_1.

We claim, however, that G_1 and G_2 delays are sufficient to recover the room in a large number of cases. First we set up the link between the geometry of the room and the measured RIR (G_1 and G_2). We assume that a genie labeled the echoes as belonging to either G_1 or G_2. In reality, we will have to do the labeling ourselves; this is discussed in Section 3.6.4.

Lemma 3.1

Let the room vertices be given in $\boldsymbol{P}$. Associate with this room a matrix $\boldsymbol{A} = \operatorname{diag}(a_1, \ldots, a_K)$ where $a_i \overset{\text{def}}{=} \|\boldsymbol{q}_i\|$ is the ith diagonal entry, and a matrix $\boldsymbol{D} \overset{\text{def}}{=} (\|\boldsymbol{q}_{ij}\|^2)$, having the second-order delays as its elements. Furthermore, let $\boldsymbol{E} = \boldsymbol{1}\boldsymbol{1}^\top$ be a $K \times K$ matrix of ones, and $\boldsymbol{N}$ be a matrix of wall normals corresponding to $\boldsymbol{P}$. Then the following holds,

$$\boldsymbol{N}^\top \boldsymbol{N} = \boldsymbol{A}^{-1}(\boldsymbol{A}^2\boldsymbol{E} + \boldsymbol{E}\boldsymbol{A}^2 - \boldsymbol{D})\boldsymbol{A}^{-1}/2. \tag{3.9}$$

Proof: From Figure 3.6 we obtain that

$$\boldsymbol{q}_i = 2\langle \boldsymbol{p}_i, \boldsymbol{n}_i \rangle \boldsymbol{n}_i. \tag{3.10}$$

In the second generation we consider each of the K first-generation virtual sources as the new source and using the same logic as above compute the second-generation virtual sources. Since in (3.10) we assumed the source to be at the origin, now we move the origin to $\boldsymbol{q}_i$,

$$\begin{aligned} \boldsymbol{q}_{ij} &= \boldsymbol{q}_i + 2\langle \boldsymbol{p}_j - \boldsymbol{q}_i, \boldsymbol{n}_j \rangle \boldsymbol{n}_j \\ &= \boldsymbol{q}_i + \boldsymbol{q}_j - 2\langle \boldsymbol{q}_i, \boldsymbol{n}_j \rangle \boldsymbol{n}_j. \end{aligned} \tag{3.11}$$

Let $a_i \stackrel{\text{def}}{=} 2\langle \boldsymbol{p}_i, \boldsymbol{n}_i \rangle = \|\boldsymbol{q}_i\|$ and $n_{ij} \stackrel{\text{def}}{=} \langle \boldsymbol{n}_i, \boldsymbol{n}_j \rangle$. Then using (3.10) and (3.11) we obtain

$$\|\boldsymbol{q}_{ij}\|^2 = a_i^2 + a_j^2 - 2a_i a_j n_{ij}. \tag{3.12}$$

This gives us the delays from G_2 in terms of the delays from G_1 and inner products between the corresponding normals. A particular consequence of (3.12) is that $\|\boldsymbol{q}_{ij}\| = \|\boldsymbol{q}_{ji}\|$. This means that in the second generation we can resolve at most $K(K-1)/2$ distinct pulses in the impulse response.

The equation (3.12) can be stated in a matrix form as

$$\boldsymbol{D} = \boldsymbol{A}^2\boldsymbol{E} + \boldsymbol{E}\boldsymbol{A}^2 - 2\boldsymbol{A}\boldsymbol{N}^\top\boldsymbol{N}\boldsymbol{A}. \tag{3.13}$$

Now the claim of the lemma follows directly. ■

Thus we get a simple expression that links the room geometry with delay times in the impulse response. Notice that a_i is the ith first generation delay, so by measuring the RIR we get access to both $\boldsymbol{A}$ and $\boldsymbol{D}$. This means that we can easily solve for $\boldsymbol{N}^\top\boldsymbol{N}$. By taking the matrix square root, we get $\boldsymbol{N}$, the matrix of normals, up to a rotation about the origin. But $\boldsymbol{A}$ and $\boldsymbol{N}$ completely determine the room shape, so estimating the room geometry becomes equivalent to an eigen-decomposition. There is a catch though: even if we know G_1 and G_2, we do not know how to order them in $\boldsymbol{A}$ and $\boldsymbol{D}$, so we end up with an assignment problem. To solve this problem we will use the properties of $\boldsymbol{D}$; we begin by a useful consequence of Lemma 3.1.

We can recognize $\boldsymbol{D}$—the matrix of distances to second-order image sources—as an EDM corresponding to the first-order image sources. This may come as a surprise, as *a priori* there is no reason to expect that the distances to second-order image sources equal the distances between the first-order image sources. We summarize it in a corollary:

Corollary 3.1

If $\boldsymbol{D}$ *is defined as in the statement of Lemma 3.1, then* $\boldsymbol{D} = \mathit{EDM}(\boldsymbol{q}_1, \ldots, \boldsymbol{q}_K)$. *Moreover,* $\operatorname{rank} \boldsymbol{D} \leqslant 4$.

Proof: That $\boldsymbol{D}$ is an EDM of $[\boldsymbol{q}_1, \ldots, \boldsymbol{q}_K]$ is immediate from the equation (3.13). We will give an alternative, visual proof in Section 3.6.4. The rank property follows directly from the fact that $\boldsymbol{D}$ is an EDM of points in $\mathbb{R}^2$ and applying Theorem 2.1. ■

3.6.2 Uniqueness of RIR: An Assignment Problem

In this section, we show that under the right circumstances it is possible to correctly assign the G_1 and G_2 delays to matrices $\boldsymbol{A}$ and $\boldsymbol{D}$, and therefore recover the room using (3.9).

Assumption 3.1

We assume that the listener at the origin hears K echoes in G_1 and $K(K-1)/2$ echoes in G_2, when an omnidirectional pulse is emitted from the origin.

We remark that this assumption can hold only in simple rooms, *e.g.*, triangles or rectangles. However, we will for now assume that we can always get these echoes, and discuss what to do about the missing ones later.

To prepare the ground for the main result, we put forward some properties of the involved matrices. Consider a K-sided room with vertices $\boldsymbol{P}$ and corresponding wall normals $\boldsymbol{N}$. Let $\boldsymbol{A}$ and $\boldsymbol{D}$ be the matrices that correspond to the room $\boldsymbol{P}$ with the correct labeling. Furthermore, let π be a permutation operator that acts on matrices in such a way that if $\boldsymbol{R}$ is some matrix related to the room $\boldsymbol{P}$ then $\pi(\boldsymbol{R})$ is the matrix that we would get if we relabeled the vertices in $\boldsymbol{P}$ according to π. Then the following lemma holds.

Lemma 3.2

Let $\boldsymbol{A}_\pi = \pi(\boldsymbol{A}), \boldsymbol{D}_\pi = \pi(\boldsymbol{D})$ and $\boldsymbol{N}_\pi^\top \boldsymbol{N}_\pi = \pi(\boldsymbol{N}^\top \boldsymbol{N})$. Then $\boldsymbol{N}_\pi^\top \boldsymbol{N}_\pi = \boldsymbol{A}_\pi^{-1}(\boldsymbol{A}_\pi^2 \boldsymbol{E} + \boldsymbol{E}\boldsymbol{A}_\pi^2 - \boldsymbol{D}_\pi)\boldsymbol{A}_\pi^{-1}/2$.

In words, we do not have to search for an absolutely correct arrangement of delays. We only have to match the permutation of $\boldsymbol{D}$ and $\boldsymbol{A}$. If we plug these into (3.9) we obtain the permuted matrix of normals $\boldsymbol{N}_\pi$. The room can be easily constructed by knowing the wall normals $\boldsymbol{n}_i$ and the distances of walls from the origin $a_i/2$. Notice that $\operatorname{rank}(\boldsymbol{N}_\pi^\top \boldsymbol{N}_\pi) = 2$.

Now we need a way to tell if a *relative* arrangement between $\boldsymbol{A}$ and $\boldsymbol{D}$ is wrong. It is important to note that we are distributing $K(K-1)/2$ elements in $\boldsymbol{D}$, and only K elements in $\boldsymbol{A}$, so the number of possible assignments is much larger in $\boldsymbol{D}$. In particular, any relabeling of $\boldsymbol{A}$ can be matched by $\boldsymbol{D}$, so we only need to look at permutations in $\boldsymbol{D}$. The results that follow will be of the almost surely type with respect to the room, so we need to assume a certain probability distribution on the set of all convex polygonal rooms. However, there is no need to specify it explicitly. All we ask is that the probability density function $f_{\boldsymbol{P}}$ and the induced distribution on the distances $(a_1, \ldots a_K)$ be absolutely continuous; most of the reasonable generative models for rooms satisfy this requirement. This discussion is formalized in the following lemma:

Lemma 3.3 (*Detectability*)

Assume that the room $\boldsymbol{P}$ is drawn according to $f_{\boldsymbol{P}}$, and consider the EDM $\boldsymbol{D}$ induced by the first-order images of the origin. Then with probability 1, the only permutations of elements in $\boldsymbol{D}$ such that the new matrix is in $\mathbb{EDM}^2$ correspond to relabelings of the points; that is, they are of the form $\boldsymbol{\Pi}\boldsymbol{D}\boldsymbol{\Pi}^\top$, where $\boldsymbol{\Pi}$ is a permutation matrix.

Proof: By Corollary 3.1 we know that $\boldsymbol{D}$ is an EDM of a set of points $\boldsymbol{Q} = [\boldsymbol{q}_1, \ldots, \boldsymbol{q}_K]$. Because of the absolute continuity of $f_{\boldsymbol{P}}$, we known by Theorem 3.1 that with probability 1, $\{\boldsymbol{q}_1, \ldots, \boldsymbol{q}_K\}$ is the only point set in $\mathbb{R}^2$ (up to a rigid transformation) that generates the set of distances in $\sqrt{\boldsymbol{D}}$. We can fix any labeling of the points, and any permutation of this labeling will lead to a distance matrix of the form $\boldsymbol{\Pi}\boldsymbol{D}\boldsymbol{\Pi}^\top$. If we rearrange the elements of $\boldsymbol{D}$ in a way that cannot be modeled as $\boldsymbol{\Pi}\boldsymbol{D}\boldsymbol{\Pi}^\top$, then the resulting matrix cannot be in $\mathbb{EDM}^2$ as this would violate the uniqueness of $\{\boldsymbol{q}_1, \ldots, \boldsymbol{q}_K\}$. ■

This means that we have a tool for detecting a wrong arrangement: First find the correct $\boldsymbol{D}$, then find the permutation of $\boldsymbol{A}$ so that $\boldsymbol{N}^\top \boldsymbol{N}$ in (3.9) have correct rank. Collecting these results, we are in a position to state the following:

Theorem 3.2

Draw a room $\boldsymbol{P}$ at random from $f_{\boldsymbol{P}}$. Then almost surely the (unlabeled) set of measurements (G_1, G_2) cannot be generated by any other room. The unique room $\boldsymbol{P}$ can be retrieved by Algorithm 3.1.

Algorithm 3.1 along with the above lemmas gives a constructive proof of the recovery of the unique room (up to a rotation and a reflection) from G_1 and G_2. While it involves a combinatorial search, it is of modest size, and various rules and heuristics can be employed to trim down the search space considerably; the most straightforward one is the triangle inequality.

3.6.3 Indoor Localization as a Byproduct

We propose several methods for indoor localization in Chapter 4. Nevertheless, we decided to mention an indoor localization method here that is an immediate byproduct of the room reconstruction algorithm.

How can we use the results of the previous section to localize a source inside a *known* or *unknown* room? The geometry estimation algorithm outputs the receiver location as a byproduct and we should clarify what is different if we know the room geometry in advance. Knowledge of the room geometry implies that each wall has an *a priori label* so the location can be given in terms of these labels (*e.g.*, 3 meters from wall A, 4 meters from wall B, and so on). This fixed labeling forces the ordering of elements in the involved matrices.

By knowing the room geometry, we know $\boldsymbol{N}$. Furthermore, we observe G_1 and G_2. By the results of the previous section, if we plug $\boldsymbol{N}^\top \boldsymbol{N}$ and $\boldsymbol{A}$ into (3.13) we should get $\boldsymbol{D}$ with the observed G_2 delays. But if we use the wrong permutation of $\boldsymbol{A}$ we end up with a wrong $\boldsymbol{D}$.

This leads to Algorithm 3.2 for indoor localization. We do not completely avoid the combinatorial search but we might be able to run this search only once in a while if we consider a scenario where the location is being updated in regular intervals. This should be possible if the displacement between two runs of the localization algorithm is not large, so that the assignment does not change between two measurements. It is easy to detect that the assignment changed, and only then we should rerun the combinatorial search.

3.6.4 A Generational Issue: An Unexpected EDM

The drawback of the mentioned algorithms is that they require the full set of echoes from G_1 and G_2. We now discuss a method that allows us to obviate this requirement.

Algorithm 3.1 Room recovery

```
(i)   Combine the delays from $G_2$ in the putative $\boldsymbol{D}$ until $\boldsymbol{D} \in \mathbb{EDM}^2$,
(ii)  Rearrange the diagonal of $\boldsymbol{A}_\pi$ until it is matched with computed $\boldsymbol{D}$, i.e., $\mathrm{rank}(\boldsymbol{A}_\pi^2 \boldsymbol{E} + \boldsymbol{E}\boldsymbol{A}_\pi^2 - \boldsymbol{D}) = 2$. If this happens for no $\boldsymbol{A}_\pi$, repeat from (i),
(iii) Find $\boldsymbol{N}_\pi$ as a matrix square root of $\boldsymbol{N}_\pi^\top \boldsymbol{N}_\pi = \boldsymbol{A}_\pi^{-1}(\boldsymbol{A}_\pi^2 \boldsymbol{E} + \boldsymbol{E}\boldsymbol{A}_\pi^2 - \boldsymbol{D}_\pi)\boldsymbol{A}_\pi^{-1}/2$,
(iv)  From the computed normals and distances to walls, reconstruct the convex polygonal room $\boldsymbol{P}$.
```

Algorithm 3.2 Localize and update

```
(i) Rearrange $\boldsymbol{A}$ until $\boldsymbol{A}^2\boldsymbol{E}+\boldsymbol{E}\boldsymbol{A}^2-2\boldsymbol{A}\boldsymbol{N}^\top\boldsymbol{N}\boldsymbol{A}$ matches the observed $G_2$ (in a noisy scenario, we look for $\boldsymbol{A}$ minimizing the $\ell_2$ distance between the delays in $G_2$ and $\boldsymbol{D}$),
(ii) To update the location, use the arrangement of $\boldsymbol{A}$ from (i) until the resulting $\boldsymbol{D}$ becomes corrupted. Additionally, since the source speed is finite we can discard all $\boldsymbol{A}$s that yield an overly large displacement,
(iii) When $\boldsymbol{D}$ becomes corrupted, repeat (i) (start by trying to swap two sides, as the displacement is still small).
```

We are already familiar with the setup in Figure 3.7 from Figure 3.6. We have added two lines to the figure that reveal a very useful property. Recall that we assume that the source and the receiver are collocated at the origin $\boldsymbol{s} = \boldsymbol{0}$. Denote by d_i the distance of the ith first-order image source to the origin (corresponding to its norm), and by d_{ij} the distance of the second-order image source to the origin.

First, notice the trapeze $(\boldsymbol{s}, \boldsymbol{q}_i, \boldsymbol{q}_{ij}, \boldsymbol{q}_j)$. The two diagonals of this trapeze must have the same length, that is,

$$\|\boldsymbol{q}_i - \boldsymbol{q}_j\| = \|\boldsymbol{q}_{ij} - \boldsymbol{0}\|. \tag{3.14}$$

In other words, the distance between the origin and the second-order image sources equals the distance between the corresponding first-order image sources,

$$d_{ij} \overset{\text{def}}{=} \|\boldsymbol{q}_{ij}\| = \|\boldsymbol{q}_i - \boldsymbol{q}_j\|, \tag{3.15}$$

which has been anticipated by the notation, and proves Theorem 3.1 visually.

The second key observation is that we can use this trapeze to construct an EDM with a special structure, whose elements are the distances to first- and second- order image sources. For example, if $K = 4$, then this EDM is given as the entrywise square of the following matrix

$$\boldsymbol{D}_{12} = \begin{array}{c} \\ \boldsymbol{s} \\ \boldsymbol{q}_1 \\ \boldsymbol{q}_2 \\ \boldsymbol{q}_3 \\ \boldsymbol{q}_4 \end{array} \begin{array}{c} \begin{array}{ccccc} \boldsymbol{s} & \boldsymbol{q}_1 & \boldsymbol{q}_2 & \boldsymbol{q}_4 & \boldsymbol{q}_4 \end{array} \\ \begin{pmatrix} 0 & d_1 & d_2 & d_3 & d_4 \\ d_1 & 0 & d_{12} & d_{13} & d_{14} \\ d_2 & d_{12} & 0 & d_{23} & d_{24} \\ d_3 & d_{13} & d_{23} & 0 & d_{34} \\ d_4 & d_{14} & d_{24} & d_{34} & 0 \end{pmatrix} \end{array}. \tag{3.16}$$

That this construction is correct follows from the definitions of d_i and d_{ij}.

The EDM $\boldsymbol{D}_{12}$ is useful because we can use matrix completion if some entries are not observed. Therefore, we do not need all second-order (or first-order) echoes, but only enough of them to determine $\boldsymbol{D}_{12}$ uniquely.

This addresses the echo unavailability problem, but $\boldsymbol{D}_{12}$ also helps with assigning echoes to the correct generation. This can again be achieved combinatorially by distributing the delays in $\boldsymbol{D}_{12}$ without knowing *a priori* to which generation they belong, and choosing the assignment that makes the putative $\boldsymbol{D}_{12}$ an EDM. We conjecture, but do not prove, that such an assignment is unique with probability one in the noiseless case. A limited number of numerical tests indicate the validity of this conjecture. The search space for the assignment can be reduced by using

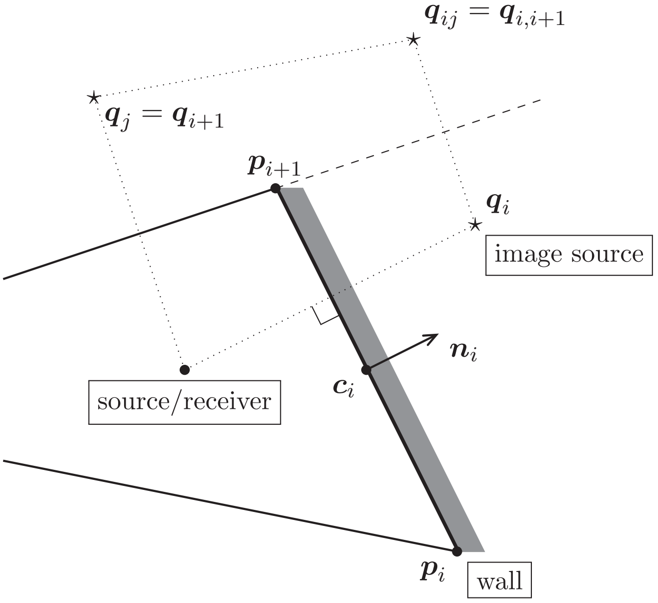


Figure 3.7: A visual proof of Corollary 3.1.

triangle inequalities such as $d_{ij} < d_i + d_j$, and the fact that $d_{ij} > d_i$ and $d_{ij} > d_j$ (*i.e.* the smallest number in each row or column must appear at the beginning).

In a K-sided room $\boldsymbol{P}$, the EDM $\boldsymbol{D}_{12}$ is uniquely determined by $2K - 1$ scalars (we need to find K points in $\mathbb{R}^2$, but we can choose the rotation of the configuration about the origin). To populate $\boldsymbol{D}_{12}$ we need $K + \frac{1}{2}K(K-1)$ delays, so the largest number of unobserved echoes is $\frac{(K-1)(K-2)}{2}$.

3.7 Numerical simulations

We validate the theoretical results on hearing the room by a single microphone by numerical simulation. We assume that the assignment of the full set of echoes to G_1 and G_2 is received as an input to the algorithm. To simulate the uncertainties in the timing estimation, we add Gaussian noise to the simulated delay times and feed them into the proposed algorithms.

Figure 3.8A shows geometry estimation for a quadrilateral room. The green line shows the estimated room in noiseless conditions and is identical to the actual room to within numerical precision. Three estimates (with different noise realizations) at a smaller jitter are plotted in red. These are barely distinguishable from the actual room shape. At a moderate jitter we observe a considerably larger deviation from the true geometry (there is one particular outlier).

A localization experiment is depicted in Figure 3.8B. The source was moving along a lemniscate with the parametric equation

$$(x, y) = \left(1 + \frac{2\cos t}{1 + \sin^2 t}, \frac{3\cos t \sin t}{1 + \sin^2 t}\right),$$

with $\Delta t = 0.1$ between localizations. As before, the green line shows the noiseless trajectory

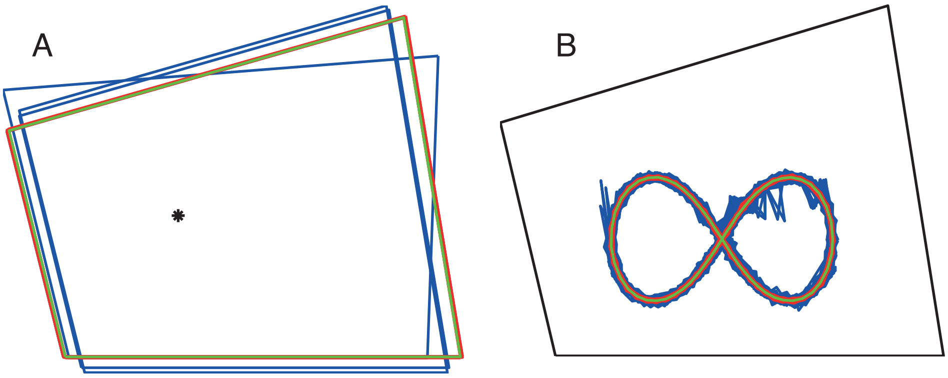


Figure 3.8: Numerical simulations with noisy delays (SNRs are indicated for the jitter added to the delays, not for the RIR signal itself). (A) Room geometry estimation, time delay SNR = 65 dB (blue), SNR = 85 dB (red) and SNR = Inf (green); (B) Source tracking: 300 realizations at SNR = 30 dB (blue), 300 realizations at SNR = 40 dB (red), noiseless (green).

tracking and is identical to the true trajectory. Estimated trajectories at two different levels of jitter are shown in red and blue.

3.8 Hearing the Room With More Than One Microphone

Combinatorial issues in separating generations of echoes from a single impulse response motivate the search for simpler and more robust solutions. In this section we discuss how to use an arbitrary microphone array in order to resolve the challenges of using only one microphone.

3.8.1 Modeling

We consider the room to be a K-faced convex polyhedron. We work in 3D, but the results extend to arbitrary dimensionalities (2D is interesting for some applications).

Same as in the case of a single microphone, we assume that the source emits a pulse and the microphones receive echoes of this pulse from which the distances between the microphones and the image sources can be computed.

In a convex room with a known source, knowing the image sources is equivalent to knowing the walls. This means that instead of searching for walls, we can search for a particular set of points. The challenge is that the distances are unlabeled: It might happen that the kth peak in the RIR from microphone 1 and the kth peak in the RIR from microphone 2 come from different walls, as illustrated in Figure 3 and Figure 4. Thus, we have to address the problem of echo labeling. The loudspeaker position need not be known; we can estimate it from the direct sound using either TOA measurements, or differences of TOAs if the loudspeaker is not synchronized with the microphones [12, 190, 119].

In practice, having a method to find good combinations of echoes is far more important than only sorting correctly selected echoes (correctly detected peaks). Impulse responses contain peaks that do not correspond to any wall. These spurious peaks can be introduced by noise, nonlinearities and other imperfections in the measurement system. We find that a good strategy is to select a number of peaks greater than the number of walls and then to prune the selection. Furthermore, some second-order echoes might arrive before some first-order ones. The image sources corresponding to second-order or higher-order echoes (see Figure 2 for an example) will be estimated as any other image source. But because we can express a second-order image source in terms of the first-order ones as

$$\widetilde{\boldsymbol{s}}_{ij} = \widetilde{\boldsymbol{s}}_i + 2\left\langle \boldsymbol{p}_j - \widetilde{\boldsymbol{s}}_i, \boldsymbol{n}_j \right\rangle \boldsymbol{n}_j, \tag{3.17}$$

and the corresponding distances as

$$\|\boldsymbol{s} - \widetilde{\boldsymbol{s}}_{ij}\| = \|\widetilde{\boldsymbol{s}}_i - \widetilde{\boldsymbol{s}}_j\|, \tag{3.18}$$

we can eliminate it during post-processing by testing the above two expressions (see Appendix 3.B.

3.8.2 Labeling the Echoes

The purpose of echo labeling is two-fold. First, it serves to remove the “ghost” echoes (that do not correspond to walls) detected at the peak-picking stage. Second, it determines the correct assignment between the remaining echoes and the walls. We propose two methods for recognizing correct echo combinations. The first one is based on the properties of Euclidean distance matrices (EDM), and the second one on a simple linear subspace condition.

3.8.3 EDM-based approach

We start by describing an approach based on EDMs to address the problem of unlabeled echoes with multiple microphones. While simpler methods can be devised to solve the problem in the noiseless case, EDMs will be the key ingredient in the noisy case.

Consider a room with a loudspeaker and an array of M microphones positioned so that they hear the first-order echoes (we typically use $M = 5$). Denote the receiver positions by $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_M$, $\boldsymbol{r}_m \in \mathbb{R}^3$ and the source position by $\boldsymbol{s} \in \mathbb{R}^3$. The described setup is illustrated in Figure 2.12.

We explain the EDM-based echo sorting with reference to this figure. Ascribe the receiver positions to the columns of $\boldsymbol{R} = [\boldsymbol{r}_1, \ldots, \boldsymbol{r}_M]$ and let $\boldsymbol{D} \in \mathbb{R}^{M \times M}$ be an EDM corresponding to the microphone setup, $\boldsymbol{D} = \mathrm{EDM}(\boldsymbol{R})$.

If the loudspeaker emits a sound, each microphone receives the direct sound and K first-order echoes corresponding to the K walls. The arrival times of the received echoes are proportional to the distances between the image sources and the microphones. As already discussed, we face a labeling problem because we do not know which wall generated which echo. This problem is illustrated in Figure 3.9 for two walls and in Figure 3.10 for the whole room. Simple heuristics, such as grouping the closest pulses or using the ordinal number of a pulse, have limited applicability, especially with larger distances between microphones. That these criteria fail is evident from the two figures.

We propose a solution based on the properties of EDMs. The loudspeaker and the microphones are—to a good approximation—points in space, so their pairwise distances form an EDM. We can exploit the rank property from Theorem 2.1: An EDM corresponding to a point set in $\mathbb{R}^n$ has rank at most $(n + 2)$. Thus in 3D, its rank is at most 5. We start from a known point set—the microphones, and attempt to add another point—an image source. This requires adding to $\boldsymbol{D}$ a row and a column which list squared distances between the microphones and the image

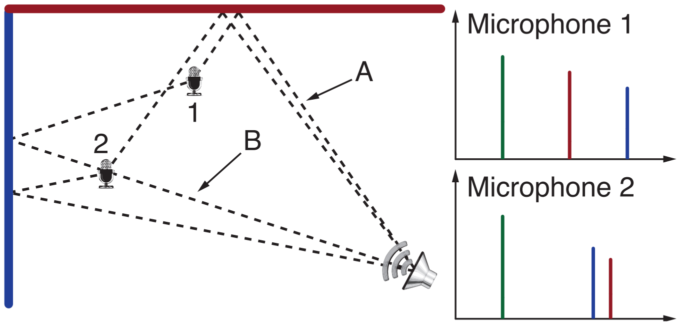


Figure 3.9: Illustration of echo swapping. Microphone 1 hears the echo from the red wall before hearing the echo from the blue wall, since path A is shorter than path B. The opposite happens for Microphone 2.

source. We extract the list of candidate distances from the RIRs, but some of them might not correspond to an image source; and for those that do correspond, we do not know to which one.

Consider again the setup in Figure 5. Microphone 1 hears echoes from all the walls, and we augment $\boldsymbol{D}$ by choosing different echo combinations. Two different augmentations are shown: $\boldsymbol{D}_{\mathsf{aug},1}$ is a plausible augmentation of $\boldsymbol{D}$ because all the distances correspond to a single image source, and they appear in the correct order. This matrix passes the rank test, or more specifically, it is an EDM. The second matrix, $\boldsymbol{D}_{\mathsf{aug},2}$, is a result of an incorrect echo assignment, as it contains entries coming from different walls. A priori, we cannot tell whether the *red* echo comes from Wall 1 or from Wall 2. It is simply an unlabeled peak in the RIR recorded by microphone 1. However, the augmented matrix $\boldsymbol{D}_{\mathsf{aug},2}$ does not pass the rank test, so we conclude that the corresponding combination of echoes is not correct.

To summarize, wrong assignments lead to augmentations of $\boldsymbol{D}$ that are not EDMs. In particular, these augmentations do not have correct rank. As it is very unlikely (as will be made precise later in Theorem 3.3) for incorrect combinations of echoes to form an EDM, we have designed a tool to detect correct echo combinations.

More formally, let $\mathcal{T}_m$ be the set of candidate distances computed from the RIR recorded by the mth microphone. We proceed by augmenting the matrix $\boldsymbol{D}$ with a combination of M unlabeled squared distances $\boldsymbol{d}$ to get $\boldsymbol{D}_{\text{aug}}$,

$$\boldsymbol{D}_{\text{aug}}(\boldsymbol{d}) = \begin{bmatrix} \boldsymbol{D} & \boldsymbol{d} \\ \boldsymbol{d}^\top & 0 \end{bmatrix}. \tag{3.19}$$

The column vector $\boldsymbol{d}$ is constructed as

$$\boldsymbol{d} = [t_1^2, \ldots, t_M^2]^\top, \tag{3.20}$$

where $t_m \in \mathcal{T}_m$.

In words, we construct a candidate combination of echoes $\boldsymbol{d}$ by selecting one echo from each microphone. We interpret $\boldsymbol{D}_{\text{aug}}$ as an object encoding a particular selection of echoes $\boldsymbol{d}$.

One might think of an EDM as a mold. It is very much like Cinderella's notorious glass slipper: If you can snugly fit a tuple of echoes in it, then they must be the right echoes. This is the key observation: If $\operatorname{rank}(\boldsymbol{D}_{\text{aug}}) < 6$ or more specifically $\boldsymbol{D}_{\text{aug}}$ verifies the EDM property, then the selected combination of echoes corresponds to an image source, or equivalently to a wall.

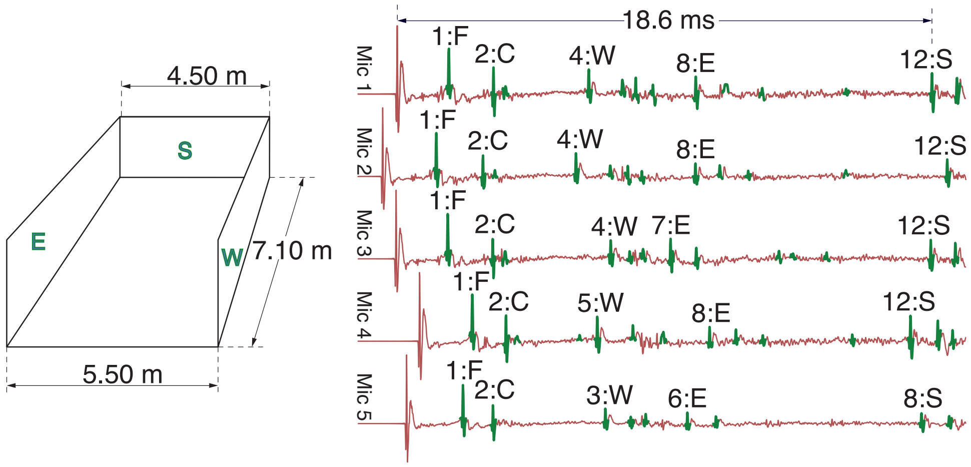


Figure 3.10: Actual room impulses responses acquired in a room sketched on the left side (see experiments for more details). First peak corresponds to direct propagation. Detected echoes are highlighted in green. Annotations above the peaks indicate the ordinal number of the peak, and the wall to which it corresponds (south, north, east, west, floor, ceiling). We can see that the ordinal number of the W-peak changes from one impulse response to another (similarly for E and S). For larger microphone arrays this effect becomes more dramatic. We also see that some peaks do not correspond to walls. Our algorithm successfully groups peaks corresponding to the same wall, and disregards irrelevant peaks.

Even if this approach requires testing all the echo combinations, in practical cases the number of combinations is small enough that this does not present a problem.

3.8.4 Subspace-based approach

An alternative method to obtain correct echo combinations is based on a simple linear condition. Note that we can always choose the origin of the coordinate system so that

$$\sum_{m=1}^{M} \boldsymbol{r}_m = \boldsymbol{0}. \tag{3.21}$$

Let $\widetilde{\boldsymbol{s}}_k$ be the location vector of the image source with respect to wall k. Then, up to a permutation, we receive at the mth microphone the squared distance information,

$$y_{k,m} \overset{\text{def}}{=} \|\widetilde{\boldsymbol{s}}_k - \boldsymbol{r}_m\|^2 = \|\widetilde{\boldsymbol{s}}_k\|^2 - 2\,\widetilde{\boldsymbol{s}}_k^\top \boldsymbol{r}_m + \|\boldsymbol{r}_m\|^2. \tag{3.22}$$

Define further $\widetilde{y}_{k,m} \overset{\text{def}}{=} -\frac{1}{2}\left(y_{k,m} - \|\boldsymbol{r}_m\|^2\right) = \boldsymbol{r}_m^\top \widetilde{\boldsymbol{s}}_k - \frac{1}{2}\|\widetilde{\boldsymbol{s}}_k\|^2$. We have in vector form

$$\begin{bmatrix} \widetilde{y}_{k,1} \\ \widetilde{y}_{k,2} \\ \vdots \\ \widetilde{y}_{k,M} \end{bmatrix} = \left[\begin{array}{c|c} \boldsymbol{r}_1^\top & -\frac{1}{2} \\ \boldsymbol{r}_2^\top & -\frac{1}{2} \\ \vdots & \vdots \\ \boldsymbol{r}_M^\top & -\frac{1}{2} \end{array}\right] \begin{bmatrix} \widetilde{\boldsymbol{s}}_k \\ \|\widetilde{\boldsymbol{s}}_k\|^2 \end{bmatrix}, \quad \text{or} \quad \widetilde{\boldsymbol{y}}_k = \boldsymbol{R}\widetilde{\boldsymbol{u}}_k. \tag{3.23}$$

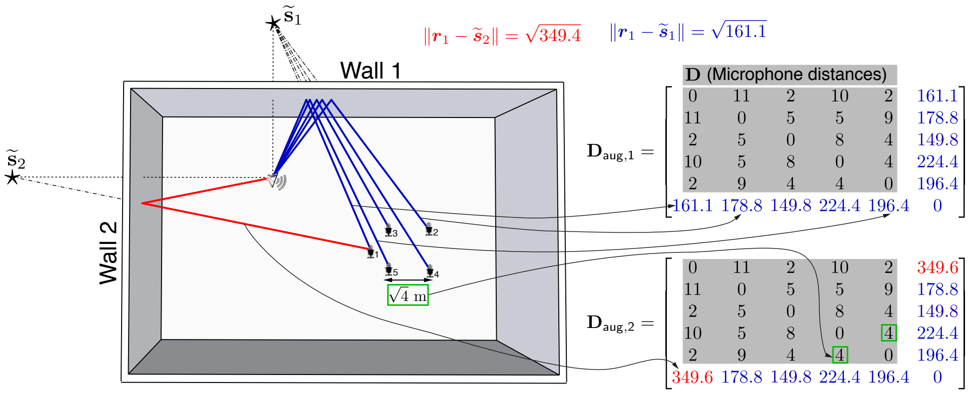


Figure 3.11: Illustration of EDM-based echo sorting. Microphones receive the echoes from all the walls, and we aim to identify echoes coming from a single wall. We select one echo from each microphone and use these echoes to augment the EDM of the microphone setup, $\boldsymbol{D}$. If all the selected echoes come from the same wall, the augmented matrix is an EDM as well. In the figure, $\boldsymbol{D}_{\text{aug},1}$ is an EDM since it contains the distances to a single point $\tilde{\boldsymbol{s}}_1$. $\boldsymbol{D}_{\text{aug},2}$ contains a wrong distance (shown in red) for microphone 1, so it is not an EDM. For aesthetic reasons the distances are specified to a single decimal place. Full precision entries are given in the SI Appendix.

Thanks to the condition (3.21), we have that

$$\mathbf{1}^\top \tilde{\boldsymbol{y}}_k = -\frac{M}{2}\|\tilde{\boldsymbol{s}}_k\|^2 \quad \text{or} \quad \|\tilde{\boldsymbol{s}}_k\|^2 = -\frac{2}{M}\sum_{m=1}^{M} \tilde{y}_{k,m}. \tag{3.24}$$

The image source is found as

$$\tilde{\boldsymbol{s}}_k = \boldsymbol{S}\tilde{\boldsymbol{y}}_k, \tag{3.25}$$

where $\boldsymbol{S}$ is a matrix satisfying

$$\boldsymbol{S}\boldsymbol{R} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}. \tag{3.26}$$

These two conditions completely characterize the distance information. In practice, it is sufficient to verify the linear constraint

$$\tilde{\boldsymbol{y}}_k \in \text{range}(\boldsymbol{R}), \tag{3.27}$$

where range($\boldsymbol{R}$) is a proper subspace when $M \geqslant 5$. But note that we can use the non-linear condition (3.24) or the condition that the matrix be in EDM^3 even if $M = 4$.

3.8.5 Uniqueness

Can we guarantee that only one room corresponds to the collected first-order echoes? To answer this, we first define the set of “good” rooms in which our algorithm can be applied. The algorithm relies on the knowledge of first-order echoes, so we require that the microphones hear them. This defines a “good” room, which is in fact a combination of the room geometry and the location of the microphone array and the loudspeaker.

Definition 3.1 (Feasibility)

Given a room $\mathcal{R}$ and a loudspeaker position $\boldsymbol{s}$, we say that the point $\boldsymbol{x} \in \mathcal{R}$ is feasible *if a microphone placed at $\boldsymbol{x}$ receives all the first-order echoes of a pulse emitted from $\boldsymbol{s}$.*

Our argument is probabilistic: The set of vectors $\boldsymbol{d}$ such that $\operatorname{rank} \boldsymbol{D}_{\text{aug}} = 5$ has measure zero in $\mathbb{R}^5$. Analogously, in the subspace formulation, $\operatorname{range}(\boldsymbol{R})$ is a proper subspace of $\mathbb{R}^5$ thus having measure zero. To use four microphones, observe for example that the same is true for the set of vectors satisfying (3.24) in $\mathbb{R}^4$. These observations, along with some technical details, enable us to state the following uniqueness result:

Theorem 3.3

Consider a room with a loudspeaker and $M \geqslant 4$ microphones placed uniformly at random inside the feasible region. Then the unlabeled set of first-order echoes uniquely specifies the room with probability 1.

The proof of the theorem is given in Appendix 3.A.

This means that we can reconstruct any convex polyhedral room if the microphones are in the feasible region. A similar result could be stated by randomizing the room instead of the microphone setup, but that would require us to go through the inconvenience of generating a random convex room. In the following, we concentrate on the EDM criterion, as it performs better in experiments.

3.8.6 Practical Algorithm

In practice, we face different sources of uncertainty. One such source is the way we measure the distances between microphones. We can try to reduce this error by calibrating the array, but we find the proposed schemes to be very stable with respect to uncertainties in array calibration. Additional sources of error are the finite sampling rate and the limited precision of peak-picking algorithms. These are partly caused by unknown loudspeaker and microphone impulse responses, and general imperfections in RIR measurement. They can be mitigated with higher sampling frequencies and more sophisticated time-of-arrival estimation algorithms. At any rate, testing the rank of $\boldsymbol{D}_{\text{aug}}$ is not a way to go in the presence of measurement uncertainties. The solution is to measure how close $\boldsymbol{D}_{\text{aug}}$ is to an EDM. We can consider different constructions:

(i) heuristics based on the singular values of $\boldsymbol{D}_{\text{aug}}$,

(ii) distance of $\widetilde{\boldsymbol{y}}_k$ from $\operatorname{range}(\boldsymbol{R})$ (equation (3.27)),

(iii) non-linear norm condition (3.24),

(iv) distance between $\boldsymbol{D}_{\text{aug}}$ and the *closest* EDM.

The approach based on the singular values of $\boldsymbol{D}_{\text{aug}}$ captures only the rank requirement on the matrix. But the requirement that $\boldsymbol{D}_{\text{aug}}$ be an EDM brings in many additional subtle dependencies between its elements. For instance, from (2.2) we have that

$$\boldsymbol{D}_{\text{aug}} \in \mathbb{EDM} \Leftrightarrow (\boldsymbol{I} - \tfrac{1}{M+1}\boldsymbol{1}\boldsymbol{1}^\top)\boldsymbol{D}_{\text{aug}}(\boldsymbol{I} - \tfrac{1}{M+1}\boldsymbol{1}\boldsymbol{1}^\top) \preceq 0. \tag{3.28}$$

Unfortunately, (3.28) does not allow us to specify the ambient dimension of the point set. Imposing this constraint leads to even more dependencies between the matrix elements, and the resulting set of matrices is no longer a cone (it is actually not convex anymore). Nevertheless, as discussed in Chapter 2, we can apply the family of algorithms used in multi-dimensional scaling (MDS) [206] to find the closest EDM between the points in a *fixed* ambient dimension.

In particular, we use the s-stress cost function (2.88) to evaluate how far the augmented "EDM" is from an actual EDM. Recall that s-stress is defined as the value of

$$\text{minimize } \|\boldsymbol{D}_{\text{aug}} - \widetilde{\boldsymbol{D}}_{\mathsf{aug}}\|_F^2 \quad \text{subject to} \quad \boldsymbol{D}_{\text{aug}} \in \mathbb{EDM}^3. \tag{3.29}$$

By $\mathbb{EDM}^3$ we denote the set of EDMs generated by point sets in $\mathbb{R}^3$. We say that s-stress is the *score* of the matrix $\widetilde{\boldsymbol{D}}_{\mathsf{aug}}$, and use it to assess the likelihood that a combination of echoes corresponds to a wall.

Reconstruction algorithm Combining the described ingredients, we design an algorithm for estimating the shape of a room. The algorithm takes as input the arrival times of echoes at different microphones (computed from RIRs). For every combination of echoes, it computes the score using the criterion of choice. We specialize to constructing the matrix $\boldsymbol{D}_{\text{aug}}$ as in (3.19) and computing the s-stress score. For the highest ranking combinations of echoes (with the smallest s-stress), it computes the image source locations. We employ an additional step to eliminate ghost echoes, higher-order image sources and other impossible solutions. Note that we do not discuss peak picking (computing the delays of the echoes).

Algorithm 3.3 summarizes the echo sorting procedure. It finds the echoes in $\mathcal{T}_2, \dots \mathcal{T}_M$ that should be combined with a particular delay t_1 received by the first microphone. The complete room shape estimation procedure can then be summarized as follows:

(i) For every echo delay $t_1 \in \mathcal{T}_1$ run Algorithm 3.3,

(ii) Compute the image source locations,

(iii) Remove image sources that do not correspond to walls (higher-order by using (3.17), "ghost" sources by heuristics),

(iv) Reconstruct the room.

Algorithm 3.3 Echo Sorting [56]

```
function EchoSort($\boldsymbol{R}, t_1, \dots, T_m$)
    $\boldsymbol{D} \leftarrow \mathrm{EDM}(\boldsymbol{R})$
    $s_{\text{best}} \leftarrow +\mathrm{Inf}$
    for all $\boldsymbol{t} = [t_2, \dots, t_m]$, such that $t_i \in T_i$ do
        $\boldsymbol{d} \leftarrow c \cdot [t_1,\ \boldsymbol{t}^\top]^\top$                ▷ $c$ is the sound speed
        $\boldsymbol{D}_{\text{aug}} \leftarrow \begin{bmatrix} \boldsymbol{D} & \boldsymbol{d} \\ \boldsymbol{d}^\top & 0 \end{bmatrix}$
        if $\mathrm{s{-}stress}(\boldsymbol{D}_{\text{aug}}) < s_{\text{best}}$ then
            $s_{\text{best}} \leftarrow \mathrm{s{-}stress}(\boldsymbol{D}_{\text{aug}})$
            $\boldsymbol{d}_{\text{best}} \leftarrow \boldsymbol{d}$
        end if
    end for
    return $\boldsymbol{d}_{\text{best}}$
end function
```

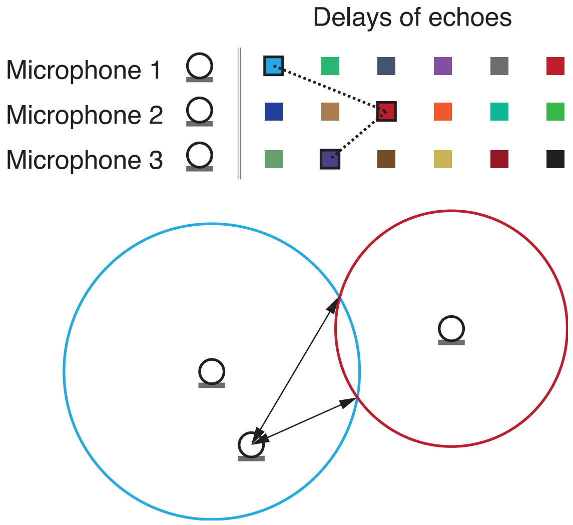


Figure 3.12: Illustration of the noiseless echo-sorting process.

Step iii) is described in more detail in Appendix 3.B. It is not necessary to test all echo combinations. An echo from a fixed wall will arrive at all the microphones within the time given by the largest inter-microphone distance. Therefore, it suffices to combine echoes within a temporal window corresponding to the array diameter. This substantially reduces the running time of the algorithm. Consequently, we can be less conservative in the peak-picking stage.

To get a ballpark idea of the number of combinations to test in the general case, suppose that we detect 20 echoes per microphone[3], and that the diameter of the five-microphone array is 1 m. Thus for every peak time $t_i \in \mathcal{T}_1$ we have to look for peaks in the remaining four microphones that arrived within a window around t_1 of length $2 \times \frac{1 \text{ m}}{343 \text{ m/s}}$, where 343 m/s is the speed of sound. This is approximately 6 ms, and in a typical room we can expect about five early echoes within a window of that duration. Thus we have to compute the s-stress for $20 \times 5^4 = 12500$ matrices of size 6×6, which can be done in a matter of seconds on a desktop computer. In fact, once we assign an echo to an image source, we can exclude it from further testing, so the number of combinations can be further reduced.

3.8.7 Computational Complexity

By employing various heuristics (based for example on the finite size of the microphone array), the described combinatorial search is reduced to a modest size. Nevertheless, we show that in the noiseless case, or with small noise, echo sorting can be performed in polynomial time.

Theorem 3.4

Assume that we have M microphones and pick K echoes per microphone. Then we can sort the echoes by testing $O(K^3)$ combinations in 2D and $O(K^4)$ combinations in 3D.

---

[3] We do not need to look beyond early echoes corresponding to at most three bounces. This is convenient, as echoes of higher orders are challenging or impossible to isolate.

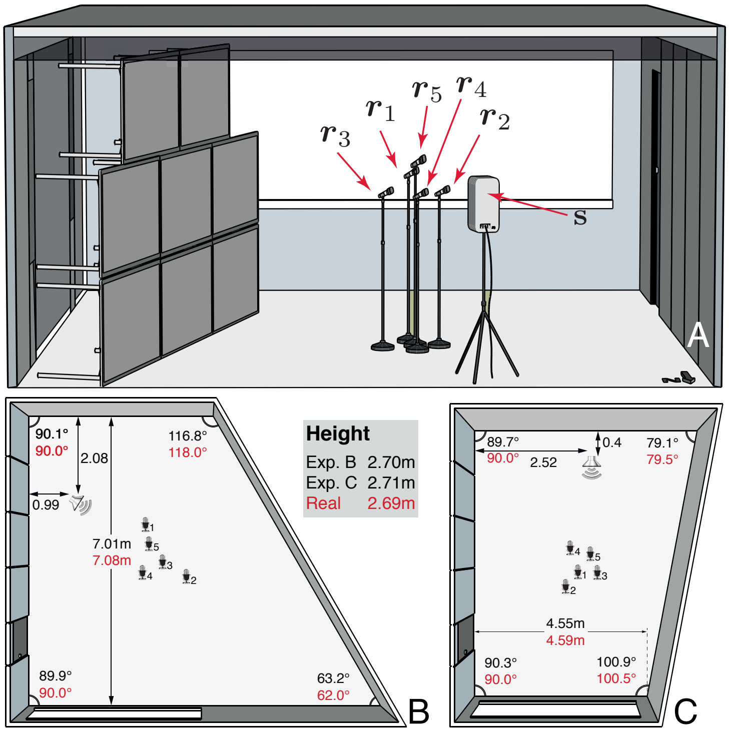


Figure 3.13: Illustration of the experiments in BC329. (A) Drawing of the room used in the experiment with a movable wall. (B, C) Two reconstruction results. Real values are indicated in red, and the estimated values are indicated in black.

Proof: We prove the theorem constructively by exhibiting an algorithm with the sought complexity. We do it in 2D, but the reasoning is identical in 3D.

Echo sorting can be described by the pictorial in Figure 3.12. It is not necessary to work with more than 3 microphones. Consider the following procedure: Pick an echo from microphone 1; it defines a circle (blue) of possible image source locations. This echo must be combined with an echo from microphone 2 and with an echo from microphone 3.

Now for every echo from microphone 2, intersect the circles from microphones 1 and 2 (subfigure B); this results in at most two intersection points. The distance from microphone 3 to one of the intersection points must appear on the list of delays for microphone 3 if we grouped echoes from microphones 1 and 2 correctly. If it does not appear on the list, we choose the next echo from microphone 3. The total number of combinations to test is then $K^2 + (K-1)^2 + \cdots + 1 = K(K+1)(2K+1)/6 = O(K^3)$. ■

Thus, without noise, or with small noise, we can run echo sorting in a time that is polynomial in the number of echoes.

3.8.8 Experiments

We ran experiments in two distinctly different environments. One set was conducted in a lecture room at EPFL, where our modeling assumptions are approximately satisfied. Another exper-

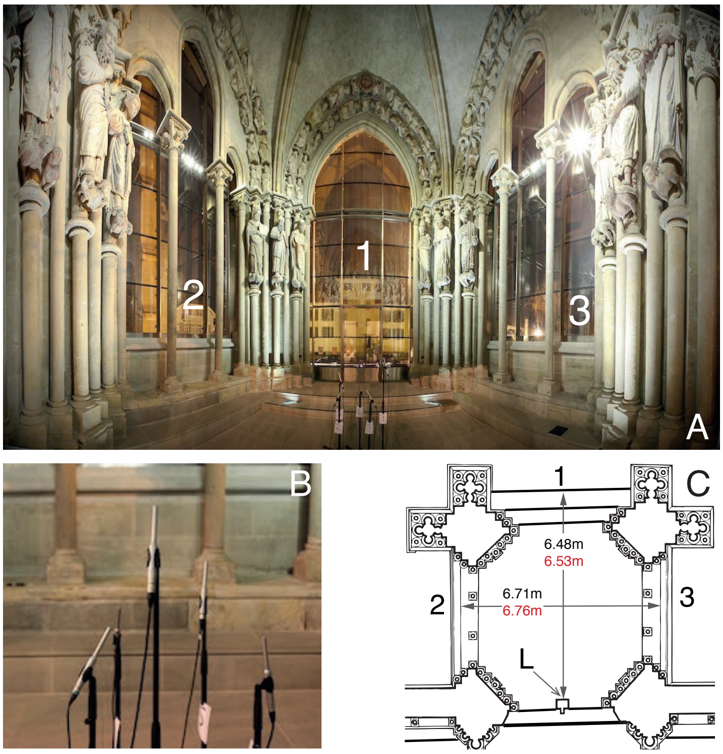


Figure 3.14: Illustration of the experiments in the Lausanne cathedral. (A) Panoramic photo of the portal of the Lausanne cathedral. (B) A close up of the microphone array used in cathedral experiments. (C) Floor plan of the portal (from [116]) with indicated measured dimensions.

iment was conducted in a portal of the Lausanne cathedral. The portal is non-convex, with numerous non-planar reflecting objects. It essentially violates the modeling assumptions, and the objective was to see whether the algorithm still gives useful information. In all experiments, microphones were arranged in an arbitrary geometry, and we measured the distances between the microphones approximately with a tape measure. We did not use any specialized equipment or microphone arrays. Nevertheless, the obtained results are remarkably accurate and robust.

The lecture room is depicted in Figure 3.13A. Two walls are glass windows, and two are gypsum-board partitions. The room is equipped with a perforated metal plate ceiling suspended below a concrete ceiling. To make the geometry of the room more interesting, we replaced one wall by tables, thus creating a movable wall. Results are shown for two positions of the table wall and two different source types. We used an off-the-shelf directional loudspeaker, an omni-directional loudspeaker and five non-matched omni-directional microphones. RIRs were estimated by the sine sweep technique [72]. In the first experiment, we used an omni-directional loudspeaker to excite the room, and the algorithm reconstructed all six walls correctly, as shown in Figure 3.13B. Note that the floor and the ceiling are estimated near-perfectly. In the second experiment, we used a directional loudspeaker. As the power radiated to the rear by this loudspeaker is small, we placed it against the north wall, thus avoiding the need to reconstruct it. Surprisingly, even though the loudspeaker is directional, the proposed algorithm reconstructs all the remaining walls accurately, including the floor and the ceiling (Figure 3.13C).

Figures 3.14A and 3.14B shows a panoramic view and the floor plan of the portal of the Lausanne cathedral. The central part is a pit reached by two stairs. The side and back walls are closed by glass windows, with their lower parts in concrete. In front of each side wall, there are two columns, and the walls are joined by column rows indicated in the figure. The ceiling is a dome approximately 9 meters high. We used a directional loudspeaker placed at the point 'L' in Figure 3.14C. Microphones were placed around the center of the portal. Alas, in this case we do not have a way to remove unwanted image sources, as the portal is poorly approximated by a convex polyhedron. The glass front, numeral 1 in Figure 3.14C, and the floor beneath the microphone array can be considered flat surfaces. For all the other boundaries of the room, this assumption does not hold. The arched roof cannot be represented by a single height-estimate. The side windows, numerals 2 and 3 in Figure 3.14C, with pillars in front of them and erratic structural elements at the height of the microphones, the rear wall, and the angled corners with large pillars and large statues, all present irregular surfaces creating diffuse reflections. In spite of the complex room structure with obstacles in front of the walls and numerous small objects resulting in many small-amplitude, temporally spread echoes, the proposed algorithm correctly groups the echoes corresponding to the three glass walls and the floor. This indicates the robustness of the method.

3.9 Conclusion

Can one hear the shape of a room? The answer is a definite *yes* in many practically relevant situations. We showed that even with a single microphone we can reconstruct the shapes of convex polyhedral rooms, at least in theory. What is more, we have shown that the convex polygonal (and polyhedral in 3D) rooms are uniquely described by G_1 and G_2, sets of times of arrivals of first- and second-order echoes. We have also described a simple indoor localization algorithm based on the proposed room reconstruction method.

What made this work is exploiting the geometrical relationships between different generations of echoes, and useful properties of Euclidean distance matrices, or more generally, of Euclidean distance geometry. This is to our knowledge the most versatile algorithm for single-channel room reconstruction in terms of reconstructible geometries, and the only one that warrants uniqueness.

Future work on single RIR room reconstruction includes

- Getting rid of the combinatorial search for assigning echoes to generations,
- Using any echo, regardless of its order,
- Studying uniqueness in more complicated geometries (curved walls, non-convex),
- Robust formulations.

Next, to address the combinatorial challenges that arise when using only a single microphone, we looked at room geometry estimation with several microphones. We presented an algorithm for reconstructing the 3D geometry of a convex polyhedral room from a few acoustic measurements. Our algorithm requires a single sound emission and uses a minimal number of microphones. The proposed algorithm has essentially no constraints on the microphone setup. Thus, we can arbitrarily reposition the microphones, as long as we know their pairwise distances (in our experiments we did not “design” the geometry of the microphone setup). Further, we proved that the first-order echoes collected by a few microphones indeed describe the room uniquely. Taking the image source point of view enabled us to derive clean criteria for echo sorting.

Our algorithms open the way for different applications in virtual reality, auralization, architectural acoustics and audio forensics. For example, we can use it to design acoustic spaces with desired characteristics or to change the auditory perception of existing spaces. The proposed echo-sorting solution is useful beyond hearing rooms. Examples are omni-directional radar, multiple-input-multiple-output (MIMO) channel estimation, and indoor localization, to name a few. As an extension of our method, a person walking around the room and talking into a cellphone could enable us to both "hear" the room and find the person's location. Future research will aim at exploring these various applications.

3.A Proof of Theorem 3.3

It is sufficient to consider $M = 4$ in 3D. For $M > 4$ we can simply select a subset of four microphones. We want to show that wrong combinations of echoes cannot produce consistent sets of distances.

A selection of echoes is a vector $\boldsymbol{d} = [t_1, \ldots, t_4]$ such that $t_m \in \mathcal{T}_m$. Notice that for any selection of echoes we must have one of the following:

(i) All echoes in $\boldsymbol{d}$ come from the same image source,

(ii) There is an image source from which exactly one echo is in $\boldsymbol{d}$,

(iii) There are two image sources such that two echoes in $\boldsymbol{d}$ come from one of them and two echoes are from the other one.

If (i) occurred, then the echoes corresponding to this image source are correctly grouped. If (ii) occurred, we can assume without the loss of generality that the image source with exactly one echo in $\boldsymbol{d}$ is S1 and that the echo is received by M1, as illustrated in Figure 3.15A. In the figure, the echoes are assumed to be sorted, so that every column corresponds to one image source. With reference to the figure, all echoes (times of arrivals) that are shaded in gray are independent from the framed blue echo in S1 by the assumptions of the theorem. A particular selection of echoes induces spheres of the radii corresponding to echo delays centered at the microphones. If the selected echoes in the shaded region induce spheres that do not intersect, we are done (this wrong combination will fail the EDM test). The echoes in the shaded part correspond to three spheres with centers at M1, M2, and M3. If all three come from the same image source, their intersection contains two points with probability one, due to the assumptions of the theorem. But then the probability that a random sphere $\|\boldsymbol{r}_1 - \boldsymbol{x}\|^2 = \|\boldsymbol{r}_1 - \boldsymbol{s}_1\|^2$ independent from these two points contains one of them is zero. Similar reasoning explains that the intersection of the three spheres in the shaded part will be either two points or an empty set with probability one.

Similarly, the case (iii) can be represented as in Figure 3.15B. Both pairs of echoes correspond to two intersected spheres, whose intersections are two circles with probability one. Due to the assumptions of the theorem, the centers of these two circles are independent and distributed absolutely continuously in 3D. Thus the probability that the two circles intersect is zero.

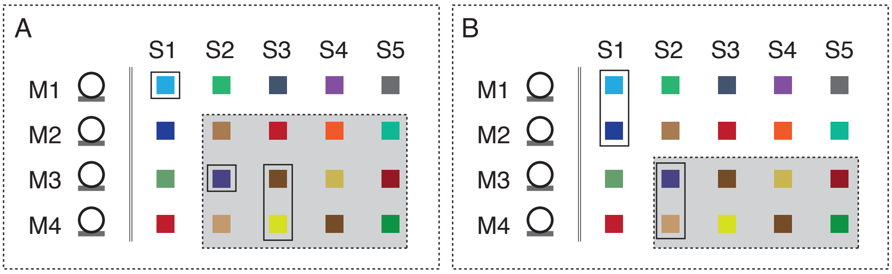


Figure 3.15: Illustration of the two salient situations in echo sorting.

3.B Room Reconstruction Procedure

The echo sorting algorithm outputs a list of image sources; some of these image sources are first-order images that we use to reconstruct the room, and some are higher-order sources. In order to reconstruct the room, we need to detect these higher-order image sources and remove them from the list. From (3.17), it follows that higher-order image sources are obtained as certain "combinations" of lower-order ones—a fact that we use to discriminate them, as explained below.

We process the candidate image sources in the order of increasing distance from the loudspeaker. If the current image source cannot be obtained as a combination of closer sources, we add the corresponding plane (halfspace) to the list of halfspaces whose intersection determines the final room.

Beyond the "combining criterion", if the halfspace (an inequality) that we are adding does not change the room, we discard the corresponding image source. We also do it if the new inequality perturbs the room only slightly.

Algorithm 3.4 Room reconstruction procedure

Input: Candidate images $\widetilde{\boldsymbol{s}}_1, \ldots, \widetilde{\boldsymbol{s}}_P$, loudspeaker location $\boldsymbol{s}_0$, distance threshold ϵ
Output: Room vertices

```
$[\widetilde{\boldsymbol{s}}_1, \cdots, \widetilde{\boldsymbol{s}}_P] \leftarrow \mathrm{SortByDistanceFromLoudspeaker}([\widetilde{\boldsymbol{s}}_1, \cdots, \widetilde{\boldsymbol{s}}_P])$
deleted$[1:P] \leftarrow$ false
for $i = 1$ to $P$ do
    if $\exists j, k < i, j \neq k$ s.t. $\|\mathrm{Combine}(\widetilde{\boldsymbol{s}}_j, \widetilde{\boldsymbol{s}}_k) - \widetilde{\boldsymbol{s}}_i\| < \epsilon$ then
        deleted$[i] \leftarrow$ true
    else if Plane$(\widetilde{\boldsymbol{s}}_i)$ intersects the current room then
        Add Plane$(\widetilde{\boldsymbol{s}}_i)$ to the current set of planes
    else
        deleted$[i] \leftarrow$ true
    end if
end for
```

This procedure is summarized in Algorithm 3.4. The following definition is used in the algorithm ($\boldsymbol{s}$ is the loudspeaker):

$$\mathrm{Combine}(\widetilde{\boldsymbol{s}}_1, \widetilde{\boldsymbol{s}}_2) \overset{\mathrm{def}}{=} \widetilde{\boldsymbol{s}}_1 + 2\langle \boldsymbol{p}_2 - \widetilde{\boldsymbol{s}}_1, \boldsymbol{n}_2 \rangle \boldsymbol{n}_2, \tag{3.30}$$

where $\boldsymbol{p}_2 = (\boldsymbol{s} + \widetilde{\boldsymbol{s}}_2)/2$ is a point on the (hypothetical) wall defined by $\boldsymbol{s}_2$; that is, a point on the median plane between the loudspeaker and $\boldsymbol{s}_2$. The outward pointing unit normal is defined as $\boldsymbol{n}_2 = (\widetilde{\boldsymbol{s}}_2 - \boldsymbol{s})/\|\widetilde{\boldsymbol{s}}_2 - \boldsymbol{s}\|$.

The room is defined as the intersection of halfspaces generated by the first-order image sources. With the above notation, halfspace corresponding to the image source $\boldsymbol{s}_i$ is defined by

$$\left\{ \boldsymbol{x} : \left\langle \boldsymbol{n}_i^\top, \boldsymbol{x} \right\rangle \leqslant \langle \boldsymbol{n}_i, \boldsymbol{p}_i \rangle \right\}. \tag{3.31}$$

The plane corresponding to the image source $\boldsymbol{s}_i$ is denoted simply by Plane$(\boldsymbol{s}_i)$.

Chapter 4

Localization and Position Calibration[*]

4.1 Introduction

In this chapter we pursue our study of the potential of echoes, this time by looking at three problems of *localization*. We will show how echoes can help localize accurately with fewer sensors and in geometries unreachable to classical methods.

A common assumption in localization methods such as beamforming, subspace methods or different parametric methods is that the sound propagates in free space [112]. These assumptions are essentially violated in rooms, and the performance of these methods degrades in the presence of multipath. The opposite is true of our methods that model the echoes: They become an additional useful source of information. As we demonstrate towards the end of the chapter, we can reap the benefits even when the geometry of the room is unknown.

We first look at the problem of localizing multiple wideband sound sources in a known room. We approach it by discretizing the Helmholtz equation using the finite element method (FEM) [194, 193], so we only consider echoes implicitly (in addition to other wave effects). FEM is especially well-suited to source localization because the obtained vectors and matrices are expressed in spatially localized basis, so we naturally obtain the measurements in the form of a sparse sum of atoms in a certain dictionary. We thus combine solving the wave equation and sparse modeling into one step, as once the geometry is specified, the physical modeling becomes implicit in the sparse approximation method.

In the second part, we move to explicit treatment of echoes which were only implicit in the first part. To use echoes effectively, we make an assumption that the source emits a known pulse whose time of arrival can be accurately measured at the receiver. This part is closely related to the topic of indoor positioning, which recently received significant attention due to numerous attractive applications. To name a few, it enables automatized inventory management, object tracking, and various location-aware services. Location-customized information is valuable for users (or information providers) in administrative buildings and museums. It is important in

---

[*] Parts of this chapter are joint work with several colleagues (all with Martin Vetterli): Single-channel localization is a joint work with Reza Parhizkar [163]; localization in non-convex rooms is a joint work with Orhan Öçal [159]; calibration using multidimensional unfolding and EDMs is a joint work with Juri Ranieri [59]; zero-knowledge calibration is a joint work with Laurent Daudet [51].

security and rescue operations, for example tracking the location of firefighters in a burning building [90].

Positioning systems (indoor or outdoor) can be divided into three main topologies [64]. First, the *self-positioning* system, where the receiver makes measurements from distributed transmitters to determine its own position (*e.g.*, GPS); second, *remote positioning*, where receivers located at possibly multiple locations measure the signal from an object to find its location; and third, *indirect positioning* where a data link is used to transfer position information from a self-positioning system to a remote site or vice versa. Our algorithms fall into first two groups, with real and image sources playing the role of distributed transmitters, and real and image microphones the role of distributed receivers.

We start by showing how to do single-channel localization. Imagine a setup where a source sends a pulse and a synchronized receiver can measure the time of arrival of the pulse. If we are to localize the source, it is clear that in free space we can do no better than to place it in on a sphere around the receiver with the radius corresponding to the propagation time. The situation changes considerably once we are in a room, as echoes provide information that can help resolve the spherical ambiguity. Next, we use echoes to localize *around corners.* Assuming that the room geometry is known, there is no need to localize the true source at all; as soon as we see *any* of the image sources, we can find the true source too.

Thirdly, we look at the problem of position calibration in microphone arrays. A compelling method to calibrate the positions of microphones in an array is with sources at unknown locations. Remarkably, it is possible to reconstruct the locations of both the sources and the receivers, if their number is larger than some prescribed minimum [186]. We propose a flexible localization algorithm by first recognizing the problem as an instance of multidimensional unfolding (MDU)—a classical problem in Euclidean geometry and psychometrics—and then solving the MDU as a special case of Euclidean distance matrix (EDM) completion. We solve the EDM completion using a semidefinite relaxation. In contrast to existing methods, the semidefinite formulation allows us to elegantly handle missing pairwise distance information, but also to incorporate various prior information about the distances between the pairs of microphones or sources. The prior information could be in the form of bounds on the distances, or it could be ordinal, such as “microphones 1 and 2 are further apart than microphones 1 and 15”. The intuition that this should improve the localization performance is confirmed by numerical experiments.

Next, we ask our usual question: Can the room help in calibration? But the case of a known room simply makes calibration equivalent to the single-channel localization scenario, so we know that the room helps. We therefore raise the stakes and ask if the room helps even if it is unknown. Thus we consider the following scenario: We place a number of microphones at unknown positions in an unknown room, and we produce a number of acoustic events at unknown positions. Can we learn the microphone locations, the room shape, and the locations of acoustic events from time of arrival measurements?

Not only is the answer affirmative, but we also show that by placing the whole setup inside an unknown room, we can reduce the number of sources required for calibration to only one. By using echoes, we can calibrate an arbitrary number of microphones in a single finger snap. Additionally, we learn the array’s absolute position inside the room; an indispensable piece of information for various tracking schemes.

It is useful to think about various inverse problems in this thesis in terms of three players in room acoustics: sources, receivers, and the room itself (abstracted through its geometry). Using this perspective, in Table 4.1 we summarize what is known and what is sought by the various algorithms.

Table 4.1: Three players in room acoustics, either known (✓) or unknown (✗) in different geometric inverse problems.

Algorithm	Sources	Receivers	Room
Room Geometry Reconstruction (Chapter 3)	✗	✓	✗
Single-Channel Localization (Section 4.3.2)	✓	✗	✓
Inverse Method of Images (Section 4.3.5)	✗	✓	✓
Zero-Knowledge Calibration (Section 4.6.2)	✗	✗	✗

4.2 Sound Source Localization with Finite Elements and Wideband Diversity

In this section we study sound source localization as a sparse reconstruction problem. Framing source localization—a parameter estimation problem—in terms of sparse reconstruction is not a new idea. Malioutov, Çetin and Willsky [133, 134] are among the first to systematically exploit spatial sparsity to localize multiple sources, and they mention several earlier references. They mainly treat narrowband sources, although some extensions to the wideband case are available in [134]. Model and Zibulevsky [149] similarly assume the sources to be sparsely distributed on a grid, but they further assume that the sources are temporally sparse in a known dictionary and show how this assumption improves the source localization performance. The idea of exploiting a known propagation model for source localization is known as *matched-field processing*, with early work in underwater acoustics already in the 1970s [27] and considerable later developments [204].

Our contributions to wave-based source localization depart from these approaches in two ways. First, we show explicitly how to discretize the wave (Helmholtz) equation in a way that is directly amenable to sparse recovery methods. This is achieved by the FEM. The key observation is that the finite support of the finite elements turns them instantaneously into a sparsifying basis. We then exploit the spatial sparsity of the sources, but assume no sparsity in the temporal domain. We assume however the knowledge of the room.

Second, if the sources are wideband, we show that by exploiting the full source bandwidth we can achieve accurate localization with fewer sensors. We improve on the treatment of the wideband case by Malioutov [134] by recognizing that the sparsity pattern does not change with the frequency, and propose greedy algorithms and algorithms based on convex optimization (group sparsity algorithms) for source localization. We recently discovered that at the time we submitted our work in September 2011 [61] to ICASSP 2012, Boufounos, Smaragdis and Raj [21] have already presented their work exploiting this constant sparsity pattern at SPIE Wavelets. Their work considered the free field, not reverberant rooms. Other works considering the wideband diversity in rooms appeared later; see for example the paper by LeRoux *et al.* [120].

All of the mentioned algorithms, including our work, exploit the sparsity of the sources in the synthesis form. They require us to compute a dictionary of Green's function, and the measurements are modeled as a sparse sum of the atoms in this dictionary. A promising alternative perspective is that of *sparse analysis*, or cosparsity, where one does not have to *a priori* invert the PDE [108, 109].

Our primary interest here is to show that rooms indeed help. To do it, we make a strong assumption that the room is known. Some recent works address a difficult problem of narrowband reverberant source localization without the knowledge of the boundary conditions. Chardon and Daudet [35, 36] propose to decompose the field into a part due to the sources (a particular solution of the Helmholtz equation without the boundary conditions), and a solution of the homogeneous equation satisfying the boundary conditions, and to model the latter as sums of plane waves or Fourier-Bessel functions. Such models are justified by the Vekua theory [150, 151, 213]. The price to pay is an expected one: considerably more sensors than when the room is known. Additionally, the antenna must enclose the region of interest.

4.2.1 Problem Setup

Consider K localized acoustic sources inside a room $\Omega \subset \mathbb{R}^3$. Assume that the spatial distribution of the sources is given by a set of points at locations $\{\boldsymbol{s}_k\}_{k=1}^K$, $\boldsymbol{s}_k \in \Omega$. The kth source's waveform is given by a signal $a_k(t)$. These signals may represent music, speech or other arbitrary sounds. The total source distribution inside the room is then described by a function f,

$$f(\boldsymbol{x}, t) = \sum_{k=1}^{K} a_k(t)\delta(\boldsymbol{x} - \boldsymbol{s}_k). \tag{4.1}$$

Sources generate pressure variations inside the room, which we denote by $u(\boldsymbol{x}, t)$. We sample $u(\boldsymbol{x}, t)$ with M microphones at positions $\{\boldsymbol{r}_m\}_{m=1}^M$ and attempt to solve the following problem:

Problem 4.1

> *Given access to measurements of sound pressure $\{u(\boldsymbol{r}_m, t) + \epsilon_m(t)\}_{m=1}^M$ inside a known room D, where $\{\epsilon_m\}_{m=1}^M$ accounts for the modeling mismatch and noise, find the source locations $\{\boldsymbol{s}_k\}_{k=1}^K$.*

4.2.2 Discretizing the Wave Equation with the Finite Element Method

Recall from Chapter 2 (*e.g.* Equation (2.24)) that the acoustic wave motion is described by the wave equation,

$$-\Delta w + \frac{1}{c^2}\frac{\partial^2 w}{\partial t^2} = g, \tag{4.2}$$

with g being the source term, not present in (2.24).

In many applications we do not require a full time-dependent wave equation, and we analyze the system under the assumption that the field is time-harmonic, $w(\boldsymbol{x}, t) = u(\boldsymbol{x}, \omega)\mathrm{e}^{-\mathrm{j}\omega t}$, which is equivalent to taking the Fourier transform of the wave equation (4.2) with respect to time. This leads to the Helmholtz equation which we have also seen in Chapter 2,

$$-\Delta u(\boldsymbol{x}, \omega) - \frac{\omega^2}{c^2} u(\boldsymbol{x}, \omega) = f(\boldsymbol{x}, \omega). \tag{4.3}$$

Equation (4.3) does not involve time derivatives since the Fourier transform simplifies them into a multiplication with the frequency squared.

Accounting for the source model (4.1), the Helmholtz equation (4.3) becomes

$$-\Delta u(\boldsymbol{x},\omega)-(\omega^2/c^2)u(\boldsymbol{x},\omega)=\sum_{k=1}^{K}a_k(\omega)\delta(\boldsymbol{x}-\boldsymbol{s}_k), \tag{4.4}$$

where s is the Fourier transform of g.

To have a complete characterization of the wave equation, we must specify the boundary conditions. Throughout this section we will assume sound-hard (perfectly reflective) walls corresponding to the Neumann boundary condition:

$$\langle\nabla u(\boldsymbol{x},t),\boldsymbol{n}(\boldsymbol{x})\rangle=0,\ \ \boldsymbol{x}\in\partial\Omega, \tag{4.5}$$

where $\boldsymbol{n}(\boldsymbol{x})$ is the unit normal, but arbitrary mixed boundary conditions are possible.

Let us show how to obtain a FEM discretization of the Helmholtz equation (4.3). We start by multiplying both sides of the equation by a *test* function v and integrating over the room,

$$-\int_\Omega \Delta u\, v\ \mathrm{d}\boldsymbol{x}\ -\ k^2\int_\Omega uv\ \mathrm{d}\boldsymbol{x}=\int_\Omega fv\ \mathrm{d}\boldsymbol{x}, \tag{4.6}$$

where $k\overset{\text{def}}{=}\omega/c$ is the wavenumber. Intuitively, if we require that this hold for any test function v, then this form is equivalent to the original *pointwise* equation. Indeed, one could imagine highly localized test functions that extract the function value at a point. Actually we require it to hold for all v that are *admissible*; admissibility conditions are discussed in detail in [194].

Equation (4.6) is asymmetric in the sense that u "has" second derivatives, while v "has" no derivatives. A more symmetric form is obtained after applying the Green's theorem (integration by parts) to the first integral,

$$\int_\Omega\langle\nabla u,\nabla v\rangle\ \mathrm{d}\boldsymbol{x}\ -\ k^2\int_\Omega uv\ \mathrm{d}\boldsymbol{x}=\int_\Omega fv\ \mathrm{d}\boldsymbol{x}. \tag{4.7}$$

Equation (4.7) is called the weak form of the Helmholtz equation. Note that the additional term (integration over the boundary $\partial\Omega$) produced by the application of Green's theorem vanishes thanks to the boundary condition (4.5).

Let us now try to find an approximate solution $u^\star\approx u$ as a linear combination of N *trial* functions $\{\phi_k\}_{k=1}^N$, $u^\star(\boldsymbol{x})=\sum_{i=1}^N u_i^\star\phi_i(\boldsymbol{x})$ by using the weak form. Plugging $u^\star$ into (4.7), we get one linear equation in N unknowns $\{u_i^\star\}_{i=1}^N$, and the problem is reduced to the computation of these coefficients. To solve it, we need N independent equations, which we can obtain by choosing N independent test functions $v_1,\ldots,v_N$.

Putting these pieces together in (4.7), the weak form becomes

$$\int_\Omega\left\langle\nabla\Big(\sum_{i=1}^N u_i^\star\phi_i\Big),\nabla v_j\right\rangle\mathrm{d}\boldsymbol{x}-k^2\int_\Omega\Big(\sum_{i=1}^N u_i^\star\phi_i\Big)v_j\ \mathrm{d}\boldsymbol{x}=\int_\Omega fv_j\ \mathrm{d}\boldsymbol{x},\quad 1\leqslant j\leqslant N. \tag{4.8}$$

For each j we have a linear equation with the unknowns $u_1^\star,\ldots,u_N^\star$. Written more compactly, the system is

$$\sum_{i=1}^N[K_{i,j}-k^2M_{i,j}]u_i^\star=f_i,\quad 1\leqslant j\leqslant N. \tag{4.9}$$

where $K_{i,j}=\int_\Omega\langle\nabla\phi_j,\nabla v_i\rangle\ \mathrm{d}\boldsymbol{x}$, $M_{i,j}=\int_\Omega\phi_i v_j\ \mathrm{d}\boldsymbol{x}$ and $f_i=\int_\Omega fv_i\ \mathrm{d}\boldsymbol{x}$, or in a matrix form,

$$\boldsymbol{u}^\star=\boldsymbol{f}. \tag{4.10}$$

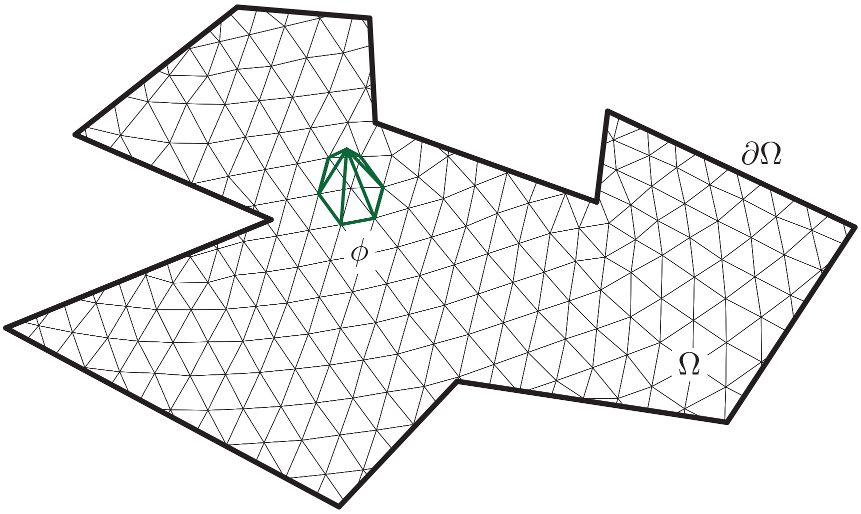


Figure 4.1: Triangular mesh in a plane. Elements ϕ are pyramids of height 1 centered at mesh nodes.

It is common to choose the test functions to be the same as the trial functions, $v_j = \phi_j$, so that $\boldsymbol{K}$ and $\boldsymbol{M}$ be symmetric positive-definite matrices.

Interestingly, the approximate solution $u^\star$ is the orthogonal projection of the exact solution u onto the linear subspace spanned by $\phi_1, \ldots, \phi_N$. This projection is called the *Galerkin projection* in honor of the Russian mathematician Boris Galerkin. From Galerkin's idea to FEM there is only one small step—one chooses the trial functions $\{\phi_i\}_{i=1}^N$ to be localized piecewise polynomials.

There are two immediate benefits from having this formulation. First, if we choose ϕs to have a localized support, many of the elements will not overlap. This means that many integrals for $K_{i,j}$ and $M_{i,j}$ will be zero, and $\boldsymbol{K} - k^2\boldsymbol{M}$ will be sparse. Second, since there are no more second derivatives, we can conveniently choose the elements ϕ_j to be piecewise linear. A typical procedure is then to discretize the domain into a triangular mesh, and use piecewise linear elements centered at the mesh nodes. An illustration in 2D is given in Figure 4.1.

We omit the discussion of how to *insert* the boundary conditions into the FEM formulation. It is an important topic though, closely connected to the uniqueness of the solution or invertibility of $\boldsymbol{K} - k^2\boldsymbol{M}$. We will simply assume that this matrix is invertible.

4.2.3 Sparsity and Wideband Diversity

An advantage of choosing finite elements for test and trial functions is their restricted spatial support. Furthermore, with piecewise linear elements, only one element at any given node is non-zero.

This implies that given a sufficiently fine mesh, the amplitude of the pressure oscillations at some location $\boldsymbol{x}$ at a frequency ω approximately equals the value of the coefficient $u^\star_{\omega,i}$ corresponding to the finite element centered around $\boldsymbol{x}$. Thus the sources of the field will *activate* only a small number of finite elements.

More concretely, let $\boldsymbol{A}(\omega) \overset{\text{def}}{=} \boldsymbol{K} - (\omega^2/c^2)\boldsymbol{M}$ be the matrix of the discretized Helmholtz equation and denote $\boldsymbol{G}(\omega) \overset{\text{def}}{=} \boldsymbol{A}(\omega)^{-1}$. Then given the source distribution $\boldsymbol{f}(\omega)$, the solution

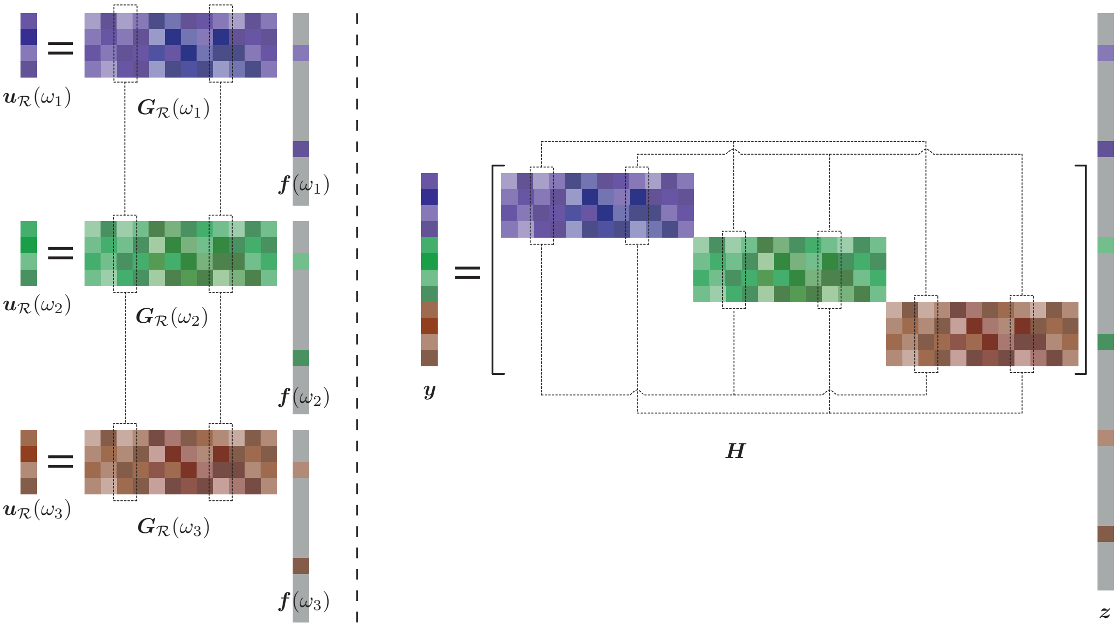

$u_{\mathcal{R}}(\omega_1)$
$G_{\mathcal{R}}(\omega_1)$
$f(\omega_1)$
$u_{\mathcal{R}}(\omega_2)$
$G_{\mathcal{R}}(\omega_2)$
$f(\omega_2)$
$u_{\mathcal{R}}(\omega_3)$
$G_{\mathcal{R}}(\omega_3)$
$f(\omega_3)$
y
H
z

Figure 4.2: Matrix structure of the wideband source localization via sparse recovery. On the left, we show systems at three different frequencies with different measurement vectors and matrices, and unknown source vectors having the same sparsity pattern. On the right, we combine the three systems into one big matrix-vector equation, with the equivalent group sparsity constraints.

$\boldsymbol{u}^\star(\omega)$ is obtained as

$$\boldsymbol{u}^\star(\omega) = \boldsymbol{G}(\omega)\boldsymbol{f}(\omega). \tag{4.11}$$

As we are solving the inverse problem, we do not know the source term $\boldsymbol{f}(\omega)$; rather, we aim to compute it from the measurements of u at a few locations. From the above discussion it is clear that by placing microphones in a room we get to measure the corresponding entries of the vector $\boldsymbol{u}^\star(\omega)$. Denote this set of entries by $\mathcal{R}$, and the restrictions of $\boldsymbol{u}^\star(\omega)$ and $\boldsymbol{G}(\omega)$ to these rows by $\boldsymbol{u}^\star_{\mathcal{R}}(\omega)$ and $\boldsymbol{G}_{\mathcal{R}}(\omega)$. Our measurements are then exactly $\boldsymbol{u}^\star(\omega)$.

Note that we implicitly assume that the microphones are located at the mesh nodes. This is not unrealistic, since we know the array geometry by design, and we can always mesh in such a way that this is true. Using the introduced notation, the measurements can be expressed as

$$\boldsymbol{u}^\star_{\mathcal{R}}(\omega) = \boldsymbol{G}_{\mathcal{R}}(\omega)\boldsymbol{f}(\omega). \tag{4.12}$$

Because the sources are spatially sparse, only a small fraction of the elements in $\boldsymbol{f}(\omega)$ are non-zero. Columns of $\boldsymbol{G}_{\mathcal{R}}(\omega)$ represent the transfer functions—samples of the Green's function—from the corresponding source location to microphones in $\mathcal{R}$. The goal is to find a sparse selection of columns in $\boldsymbol{G}_{\mathcal{R}}(\omega)$ that explains the measurements $\boldsymbol{u}^\star_{\mathcal{R}}(\omega)$.

So far we discussed only what happens at a single frequency ω. The key insight is that we can write out the Helmholtz equation for more than one ω. If the sources are wideband and their spectra are non-zero at this frequency, the solution u will be non-zero. In the inverse problem that we are solving, we will have a different measurement vector $\boldsymbol{u}^\star_{\mathcal{R}}(\omega)$ and a different sampling matrix $\boldsymbol{G}_{\mathcal{R}}(\omega)$ for every ω, but the sparsity pattern of the sought vector $\boldsymbol{f}(\omega)$ will remain unchanged, as the sources are at the same locations regardless of the frequency.

By choosing a discrete set of frequencies $\{\omega_i\}_{i=1}^F$, we arrive at the following system of underdetermined linear systems linked by a common sparsity pattern,

$$\begin{cases} \boldsymbol{u}^{\star}_{\mathcal{R}}(\omega_1) & = \boldsymbol{G}_{\mathcal{R}}(\omega_1)\boldsymbol{f}(\omega_1) \\ \boldsymbol{u}^{\star}_{\mathcal{R}}(\omega_2) & = \boldsymbol{G}_{\mathcal{R}}(\omega_2)\boldsymbol{f}(\omega_2) \\ \quad\vdots & \\ \boldsymbol{u}^{\star}_{\mathcal{R}}(\omega_F) & = \boldsymbol{G}_{\mathcal{R}}(\omega_F)\boldsymbol{f}(\omega_F) \end{cases}, \tag{4.13}$$

The structure of this system is illustrated on the left-hand side of Figure 4.2 for $F = 3$. The right hand side of the figure illustrates a way to concisely write this system in one matrix-vector equation. Let $\boldsymbol{y}^\top \overset{\text{def}}{=} [\boldsymbol{u}^{\star}_{\mathcal{R}}(\omega_1)^\top, \ \ldots, \ \boldsymbol{u}^{\star}_{\mathcal{R}}(\omega_F)^\top]$, $\boldsymbol{H} \overset{\text{def}}{=} \text{diag}(\boldsymbol{G}_{\mathcal{R}}(\omega_1), \ \ldots, \ \boldsymbol{G}_{\mathcal{R}}(\omega_F))$ and $\boldsymbol{z}^\top \overset{\text{def}}{=} [\boldsymbol{f}(\omega_1)^\top, \ \ldots, \ \boldsymbol{f}(\omega_F)^\top]$. We then rewrite (4.13) simply as

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{z}. \tag{4.14}$$

This problem is a problem of joint sparsity. Define the group of entries corresponding to the ith position in the room as

$$\boldsymbol{z}_i \overset{\text{def}}{=} \begin{bmatrix} f_i(\omega_1) \\ \vdots \\ f_i(\omega_F) \end{bmatrix}. \tag{4.15}$$

Then our goal can be stated as minimizing the number of active (non-zero) groups, while reproducing the measurements (satisfying (4.14)) as accurately as possible. Minimizing the number of active groups corresponds to minimizing the number of active sources; what happens within each group does not matter. For example, we do not require the sources to have a sparse spectrum. This structure, often termed *structured*, *joint*, *block* or *group* sparsity [10, 70, 208] is known to be promoted by the mixed ℓ^1/ℓ^2 norm[1], which is the ℓ^1 norm of the ℓ^2 norms of groups. Thus we can estimate the source locations by solving the following convex optimization program,

$$\begin{aligned} &\underset{\boldsymbol{G}}{\text{minimize}} \quad \|[\boldsymbol{z}_1, \ \ldots, \ \boldsymbol{z}_N]\|_{\ell^1/\ell^2} \ \ \Big(= \sum_{i=1}^N \|\boldsymbol{z}_i\|_2 \Big) \\ &\text{subject to} \quad \boldsymbol{y} = \boldsymbol{H}\boldsymbol{z}. \end{aligned} \tag{4.16}$$

Depending on the problem dimensions (number of possible source locations N, number of microphones M, number of frequencies F), solving the convex program (4.16) exactly may be computationally too expensive. Iterative and greedy approaches can then be used to reduce the computational cost. In [61] we used an adaptation of the orthogonal matching pursuit (OMP) [166, 209]. The performance is improved with respect to block OMP of Eldar, Kuppinger and Bölsckei [70] by simply replacing an incoherent sum of the *goodness* of groups by a coherent one. Boufounos, Smaragdis and Raj [21] propose to use the multichannel extension of the CoSaMP algorithm [156]. The modified CoSaMP requires a smaller number of measurements than the block OMP for the same sparsity level, at a somewhat higher computational cost. The best performing procedure is to solve the convex program (4.16). Knowing that the ℓ^1/ℓ^2 norm has

---

[1] Actually we want to minimize the ℓ^0/ℓ^2 norm, where ℓ^0 counts the number of non-zeros in a vector. However, this is not computationally tractable, so we do a standard trick of replacing it by its convex envelope ℓ^1/ℓ^2.

an efficient proximal mapping [192] (*cf.* Chapter 7), this minimization can also be implemented efficiently by various proximal algorithms [164].

4.2.4 Using Only One Microphone

A remarkable thing is possible if we solve the full convex relaxation (4.16) instead of using the mentioned greedy algorithms: We can localize multiple sources using a single microphone. Every pair of microphone-source locations in a room exhibits a particular spectral pattern. This pattern should be seen as a filter which is with high probability unique to this pair of locations. If a source excites this spectral pattern at the receiver, we may be able to recognize it. A few sources will excite a sparse combination of spectral patterns at the receiver, and we might be able to recover their locations from this combination.

Of course, one could imagine sources that are *adversarial* with respect to any given room, in the sense that their spectrum inverts the spectral signature of their true location and produces the signature of a different location, thus preventing correct localization. But numerical experiments reveal that the localization of multiple sources with one microphone is possible for various signals, so this is very unlikely to happen over wider frequency ranges and for various natural, noise-like signals, so our method will still work. The explanations is that inverting the frequency response is too costly in the sense that it generates a spectral signature with a large norm, thereby disqualifying the corresponding solution candidate.

What is more, if we are indeed searching for the *sparsest* combination, then the original combination will very likely be there. Again we could imagine an adversarial pair of sources jointly producing the signature of some third location at the sensor, but this would again require the costly inversion.

There is no magic here: As Boufounos notes [21], if the number of sources is greater than the number of microphones, we cannot reconstruct the source spectra as the system is underdetermined even when restricted only to active locations. To deal with this issue, we could imagine adding side knowledge via additional regularizations beyond the standard ℓ^1/ℓ^2.

As a final note, single microphone localization of multiple sources is challenging (if at all possible) in free field (see, *e.g.*, the setup in [21]) as the corresponding spectral signatures are just simple delays and thus not rich enough. It also cannot be achieved by algorithms such as CoSaMP or OMP (without additional modifications), as they do not discriminate based on spectral patterns. Concretely, the location candidates in CoSaMP are chosen iteratively as the large ℓ^2-norm rows in the following matrix:

$$\sum_{i=1}^{F} \boldsymbol{G}_{\mathcal{R}}(\omega_i)^\top \boldsymbol{u}_{\mathcal{R}}^\star(\omega_i). \tag{4.17}$$

But when $M = 1$, the above matrix-vector multiplications become scalings of a column vector for every frequency, and there is no particular selectivity for any location as all of them are scaled equally (this is not the case for $M > 1$, when the correlations between the microphone measurements and the residual vector in fact do exhibit spatial selectivity).

4.2.5 Numerical Experiments

We have validated the theoretical results in a number of numerical simulations. For visualization we use a 2D room, but the developed theory is valid in both 2D and 3D.

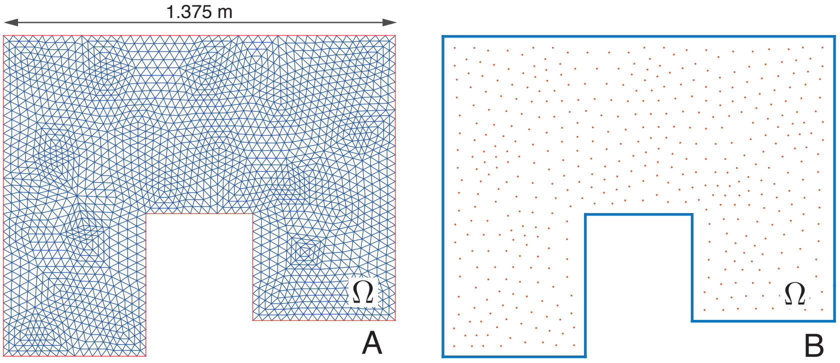


Figure 4.3: (A) Simulation domain (*room*) and the FEM simulation mesh used in the experiments. (B) Candidate source locations.

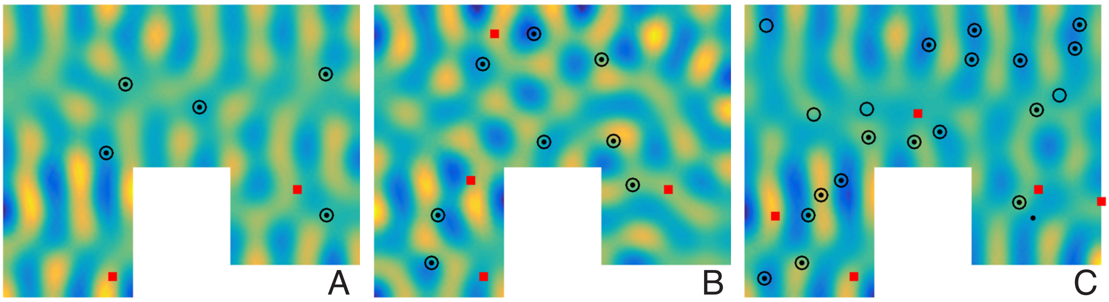


Figure 4.4: Three localization examples for $F = 512$ wavenumbers evenly spaced in $[10, 70]$ m^{-1} and source emitting white noise. Circles represent true source locations, dots represent estimated source locations, and red squares represent microphone locations. (A) 5 sources localized with 2 microphones; (B) 8 sources localized with 4 microphones; (C) attempt to localize 20 sources with 5 microphones, with erroneous localizations. Colors depict the field at the wavenumber $k = 33.37$ m^{-1}.

Acoustics are simulated using a fine mesh shown in Figure 4.3A for $F = 512$ wavenumbers k uniformly spaced in [10, 70] m^{-1}, and this result is considered to be the true acoustic wavefield. We assume that the sources can appear at a discrete set of locations shown in Figure 4.3B. In order to test the method "in isolation", we commit an inverse crime by using the same set of points to look for the sources. Nevertheless, we experiment with adding noise to the measurements (Figure 4.5).

Three localization results are shown in Figure 4.4, all with a higher number of sources than microphones. We can conclude that the room indeed helps: The (strong) assumption that we know the room geometry turns it into a highly discriminative device for source localization. The key here is to use the diversity offered by the bandwidth, thus increasing the *effective* number of measurements.

Finally, Figure 4.5 shows a localization experiment with a single microphone. We see that using the mixed-norm minimization (4.16) we can localize multiple sources using only a single sensor. What is more, the source need not be *visible* to the receiver. Figure 4.5C shows the

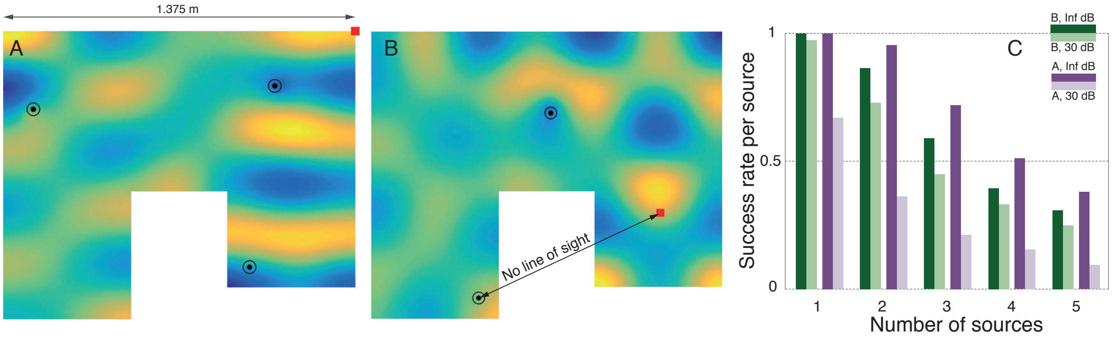


Figure 4.5: Single microphone experiments, with the same annotations as in Figure 4.4. (A) 3 sources localized by one microphone; (B) 2 source localized by one microphone, with one source being *invisible* to the microphone; (C) success rate per source for two different microphone locations, and the number of source K between 1 and 5. Colors depict the field at $k = 15.8$ m^{-1} in (A) and at $k = 22.8$ m^{-1} in (B).

localization success rate for different numbers of randomly placed sources and two different microphone locations. The success rates are shown both for the noiseless case and at the SNR of 30 dB. Two observations can be made: 1) microphone location matters, 2) the performance is maintained in the presence of noise.

4.3 Distance-Based Approaches (or a Tale of Two Applications of Echo Sorting)

In this section we present two simple localization results that exploit echoes and echo sorting. Together with Section 4.4 they may be seen as setting up the stage for the result on zero-knowledge calibration. We start by showing that in a room, it is possible to localize an omnidirectional source with a single omnidirectional receiver (and vice versa).

Localizing a source usually involves more than one receiver, especially if the receivers are omnidirectional [12, 135]. We show how to use only one sensor for source localization in a room by explicitly taking echoes into account [163]. To do it, we will assume that the room geometry is known. We do not require the detailed knowledge of the room structure—locations and orientations of principal reflectors (walls) are sufficient. The key ingredient is to recognize that the image sources and image microphones provide additional measurements *for free*. In the context of multilateration (Section 4.3.1) these additional sources or sensors are often called *virtual anchors*.

Note that we can either localize sources using multiple receivers, or localize receivers using multiple sources. In the former case we would have virtual microphones that behave analogously to the virtual sources by the duality of the wave equation. For the sake of consistency (that is, to keep working with image sources), we first consider the case of microphone localization using multiple image sources.

Virtual anchors have been used previously for similar tasks. Ribeiro *et al.* [175, 177] use echoes to improve the accuracy of source localization with a microphone array. Meissner, Steiner, and

Witrisal [143, 144] use virtual anchors for ultra-wideband (UWB) positioning, also using a single sensor. They develop a probabilistic method for localization under echo ordering ambiguity, and the source location is obtained by computing a MAP estimate. Their formulation allows them to handle missing echoes. Echo sorting, on the other hand, solves the problem in a clear-cut way with very low complexity, and gives necessary and and sufficient conditions for the localization to succeed.

Our method can be regarded as a dual of the room geometry reconstruction algorithm. Similarly to room reconstruction, we need to use echo sorting to associate recorded echoes to correct walls. In a noisy situation we can again use EDM for echo sorting, but we opt to present a simple method based on multilateration. We verify the method by an experiment in a classroom at EPFL.

4.3.1 Multilateration

At the core of our algorithms lies the principle of multilateration [12, 135]. Suppose that we have M sensors at locations $\boldsymbol{r}_m$ and a source at location $\boldsymbol{s}$. Furthermore, we have a collection of noisy measurements of the distances between $\boldsymbol{r}_m$ and $\boldsymbol{s}$,

$$d_m = \|\boldsymbol{r}_m - \boldsymbol{s}\| + \epsilon_m. \tag{4.18}$$

Multilateration refers to any method that attempts to find the source locations $\boldsymbol{s}$ given these measurements and the microphone locations.

A natural way to proceed is to find $\boldsymbol{s}$ as a minimizer of a particular cost function:

$$\widehat{\boldsymbol{s}} = \arg\min_{\boldsymbol{x}\in\mathbb{R}^3} \sum_{m=1}^{M} \big(\|\boldsymbol{r}_m - \boldsymbol{x}\| - d_m\big)^2. \tag{4.19}$$

As a motivation for this cost function, we can take the following proposition, the proof of which is given in Appendix 4.A:

Proposition 4.1

Suppose that ϵ_m are iid zero-mean Gaussian random variables. Then (4.19) *is a maximum likelihood estimate of $\boldsymbol{s}$.*

While it is desirable to compute $\widehat{\boldsymbol{s}}$ according to (4.19), the cost function is not convex and there is no known efficient algorithm for minimizing it. There exist approaches using semidefinite relaxations [37]; however, they are not guaranteed to produce the optimal solution.

Another reasonable cost function is a slight modification of (4.19), involving squared distances instead of distances. The source location estimate is thus obtained as follows:

$$\widehat{\boldsymbol{s}} = \arg\min_{\boldsymbol{x}\in\mathbb{R}^3} \sum_{m=1}^{M} \big(\|\boldsymbol{r}_m - \boldsymbol{x}\|^2 - d_m^2\big)^2. \tag{4.20}$$

This function can be seen as s-stress (*cf.* Chapter 2, Equation (2.88)) written out for a single point $\boldsymbol{x}$. Although (4.20) is also not convex, the globally optimal solution can be found efficiently (in polynomial time) using an algorithm by Beck, Stoica and Li [12]. An example of multilateration using the two cost functions is shown in Figure 4.6. Beck, Stoica and Li also devise an efficient procedure for multilateration from *distance difference* data. This corresponds to the situation where the source and the receivers are not synchronized. All of the results in this chapter can be recast in the language of distance differences, but for simplicity we choose to work with distances.

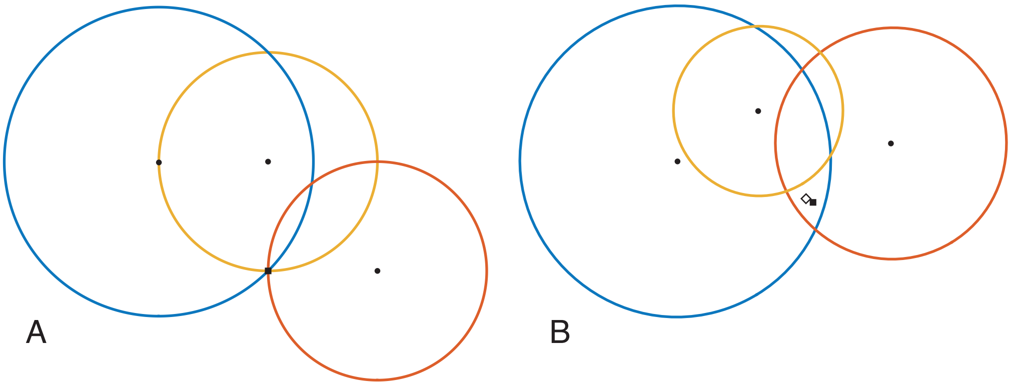


Figure 4.6: Localization by minimization of ℓ^2 error in distances squared and ℓ^2 error in non-squared distances. (A) noiseless case; (B) noisy case, the hollow diamond corresponds to distances, the full square corresponds to distances squared.

4.3.2 Single-Channel Localization

Consider now the case where $M = K = 1$, and the source emits a single click. In free space this gives us a single distance measurement d given that the source and the receiver are synchronized. Once the inflating sphere of sound passes the microphone, it is lost forever; there are no echoes that could bring it back. The equation that captures this scenario is

$$\|\boldsymbol{s} - \boldsymbol{r}\| = d, \tag{4.21}$$

which determines $\boldsymbol{s}$ up to a sphere of ambiguity if $\boldsymbol{r}$ is known, and vice-versa.

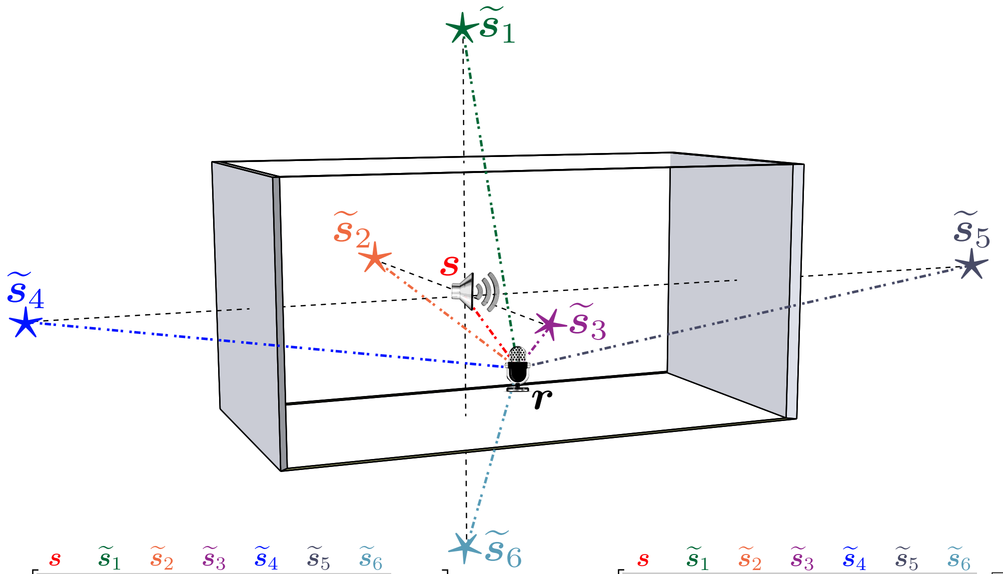
Now consider the same setup in a known room (a K-faced polyhedron). We already know that by measuring the room impulse response we access the echo arrival times t_i. These arrival times can be linked to the room geometry and microphone position with the image source model (*cf.* Chapter 2). On the other hand, because we know the room shape and the source locations, we also know the locations of the image sources. This means that instead of having just one TOA measurement, we get one per image source. This is illustrated in Figure 4.7.

As the geometry of the room and the position of the loudspeaker are known, we also know the positions of image sources, $\widetilde{\boldsymbol{s}}_i$ (*cf.* (2.62)). If we further knew the correspondences of the recorded echoes by the microphone with the image sources (which echo comes from which wall) we could simply multilaterate the position of the microphone. However, we face two problems:

- Not all the extracted echoes from the impulse response correspond to first-order image sources, and some spurious peaks do not correspond to echoes at all,

- The echoes arrive at the microphone in an unknown order that depends on the position of the microphone, so we do not know *which circle to put around which image source.*

Recall the annotated RIR measurements from Figure 3.10. We can see that many of the extracted peaks do not correspond to a first-order image source, or that they are simply spurious peaks in the impulse response. We again face an echo sorting problem.


$\widetilde{\boldsymbol{s}}_1$
$\widetilde{\boldsymbol{s}}_2$
$\boldsymbol{s}$
$\widetilde{\boldsymbol{s}}_3$
$\widetilde{\boldsymbol{s}}_4$
$\widetilde{\boldsymbol{s}}_5$
$\boldsymbol{r}$
$\widetilde{\boldsymbol{s}}_6$

$$\boldsymbol{D}_{\mathsf{aug},1} = \left[\begin{array}{cccccccc}
\boldsymbol{s} & \widetilde{\boldsymbol{s}}_1 & \widetilde{\boldsymbol{s}}_2 & \widetilde{\boldsymbol{s}}_3 & \widetilde{\boldsymbol{s}}_4 & \widetilde{\boldsymbol{s}}_5 & \widetilde{\boldsymbol{s}}_6 & \\
0.00 & 1.00 & 1.44 & 0.64 & 2.56 & 5.76 & 1.00 & 0.12 \\
1.00 & 0.00 & 2.44 & 1.64 & 3.56 & 6.76 & 4.00 & 1.52 \\
1.44 & 2.44 & 0.00 & 4.00 & 4.00 & 7.20 & 2.44 & 2.04 \\
0.64 & 1.64 & 4.00 & 0.00 & 3.20 & 6.40 & 1.64 & 0.44 \\
2.56 & 3.56 & 4.00 & 3.20 & 0.00 & 16.0 & 3.56 & 3.32 \\
5.76 & 6.76 & 7.20 & 6.40 & 16.0 & 0.00 & 6.76 & 4.92 \\
1.00 & 4.00 & 2.44 & 1.64 & 3.56 & 6.76 & 0.00 & 0.72 \\
0.12 & 1.52 & 2.04 & 0.44 & 3.32 & 4.92 & 0.72 & 0.00
\end{array}\right]
\qquad
\boldsymbol{D}_{\mathsf{aug},2} = \left[\begin{array}{cccccccc}
\boldsymbol{s} & \widetilde{\boldsymbol{s}}_1 & \widetilde{\boldsymbol{s}}_2 & \widetilde{\boldsymbol{s}}_3 & \widetilde{\boldsymbol{s}}_4 & \widetilde{\boldsymbol{s}}_5 & \widetilde{\boldsymbol{s}}_6 & \\
0.00 & 1.00 & 1.44 & 0.64 & 2.56 & 5.76 & 1.00 & \boxed{4.92} \\
1.00 & 0.00 & 2.44 & 1.64 & 3.56 & 6.76 & 4.00 & 1.52 \\
1.44 & 2.44 & 0.00 & 4.00 & 4.00 & 7.20 & 2.44 & 2.04 \\
0.64 & 1.64 & 4.00 & 0.00 & 3.20 & 6.40 & 1.64 & 0.44 \\
2.56 & 3.56 & 4.00 & 3.20 & 0.00 & 16.0 & 3.56 & 3.32 \\
5.76 & 6.76 & 7.20 & 6.40 & 16.0 & 0.00 & 6.76 & \boxed{0.12} \\
1.00 & 4.00 & 2.44 & 1.64 & 3.56 & 6.76 & 0.00 & 0.72 \\
\boxed{4.92} & 1.52 & 2.04 & 0.44 & 3.32 & \boxed{0.12} & 0.72 & 0.00
\end{array}\right]$$

Figure 4.7: An example of echo labeling for microphone localization. The gray part of the matrices shows the distances between the sources. We augment this matrix with a combination of echoes extracted from the microphone RIR. If the echoes are selected correctly and have the right order the augmented matrix is an EDM. The matrix $\boldsymbol{D}_{\mathsf{aug},1}$ is an EDM. But since the echoes are not correctly ordered in $\boldsymbol{D}_{\mathsf{aug},2}$, it is not an EDM.

4.3.3 Echo Labeling and Almost Sure Uniqueness

To correctly label the echoes, we could use the same EDM-based algorithm we used to sort echoes in the room hearing application in Chapter 3. However, let us explain a similar (but not equivalent) procedure in terms of multilateration.

Denote by $d_1, \ldots, d_M$ the measured distances between the sources (real or virutal) $\boldsymbol{s}_k$ and the microphone whose position $\boldsymbol{r}$ we seek to estimate, and define f as follows:

$$f(d_1, \ldots, d_M) = \min \left\{ \sum_{k=1}^{K} \left(\|\boldsymbol{s}_k - \boldsymbol{x}\|^2 - d_m^2 \right)^2 \;\middle|\; \boldsymbol{x} \in \mathbb{R}^3 \right\}. \tag{4.22}$$

That is, f is the optimal value of the multilateration cost function with squared distances (4.20), and we know that it can be computed efficiently [12]. In the noiseless case we are searching for an assignment of echoes that makes f zero; in the noisy case, we want to make f as small as possible.

Let $\mathcal{T}$ be the set of extracted times of arrivals of echoes (including the direct sound), and let $[\boldsymbol{s}_k]_{k=1}^{K}$ be the list of all anchors (real and virtual). Then we define an *echo assignment* as follows,

Definition 4.1 (*Echo assignment*)

An echo assignment is a K-tuple $(d_1, \ldots, d_K)$ such that $d_k \in \mathcal{T}$ and $d_k \neq d_\ell$ for $k \neq \ell$. The set of all assignments is denoted by $\mathcal{A}(\mathcal{T})$.

To model situations with multiple echoes arriving at the same time, we can extend $\mathcal{T}$ to be a multi-set, and relax the requirement that d_k be distinct. We choose not to pursue this possibility here.

Drawing on the discussions in Chapter 3, we know that with high probability, only the correct selection of echoes can yield a zero cost function in the noiseless case, and that in a noisy situation we should expect the correct combination to yield the smallest value of the cost function. Thus the correct labeling can be obtained as follows,

$$(\widehat{d}_1, \ldots, \widehat{d}_M) = \underset{(d_1,\ldots,d_K)\in\mathcal{A}(\mathcal{T})}{\arg\min} f(d_1, \ldots, d_M). \tag{4.23}$$

Using familiar arguments, we can establish the following proposition:

Proposition 4.2

Suppose that we detect L echoes (including the direct path) in the received signal and thus obtain the corresponding image source distances $\mathcal{T} = \{\|\boldsymbol{r} - \boldsymbol{s}_\ell\|\}$. Then the noiseless single-channel source localization problem can be solved by testing $O(L^3)$ echo assignments of size $K = 4$.

The above procedure is similar to using EDMs as in echo sorting (*cf.* Algorithm 3.3), but not identical. The principal difference is that here we implicitly assume that the locations of the microphone and virtual microphones are perfectly known. In reality we also want to optimize over their locations so it makes sense to run everything EDM style (although we could first *perfect* our knowledge of the virtual sources or microphones).

Using the techniques introduced in Chapter 3, we can further establish the following uniqueness result:

Proposition 4.3

Consider any full-dimensional convex polyhedral room Ω, with wall normals $\boldsymbol{n}_i$ and $\boldsymbol{p}_i$ being any point belonging to the ith wall, and place a source $\boldsymbol{s}$ and a receiver $\boldsymbol{r}$ in the room independently at random according to some absolutely continuous distribution. Then the location of $\boldsymbol{r}$ is uniquely determined by any (unlabeled) subset of $\big\{\|\boldsymbol{r} - \boldsymbol{s}_i\| \;\big|\; \boldsymbol{s}_i = \boldsymbol{s} + \langle \boldsymbol{p}_i - \boldsymbol{s}, \boldsymbol{n}_i \rangle \boldsymbol{n}_i \big\}$ of cardinality at least four.

Note that there is no reason not to consider image sources of arbitrarily high orders, but in practice high-order echoes are difficult to measure.

4.3.4 Experimental Example

We ran an experiment in a lecture room on EFPL campus (the same room we used for experiments in Chapter 3). Two walls of the room are glass windows, and two are gypsum-board partitions. The room is equipped with a perforated metal plate ceiling suspended below a concrete ceiling. We replaced one wall by a wall made of tables. The RIR from the loudspeaker to the microphone was estimated by the sine sweep technique [72]. The room dimensions are known *a priori* and the loudspeaker location was measured during the experiment. The experimental setup with the image sources of the loudspeaker are shown in Figure 4.8. As the loudspeaker is placed against

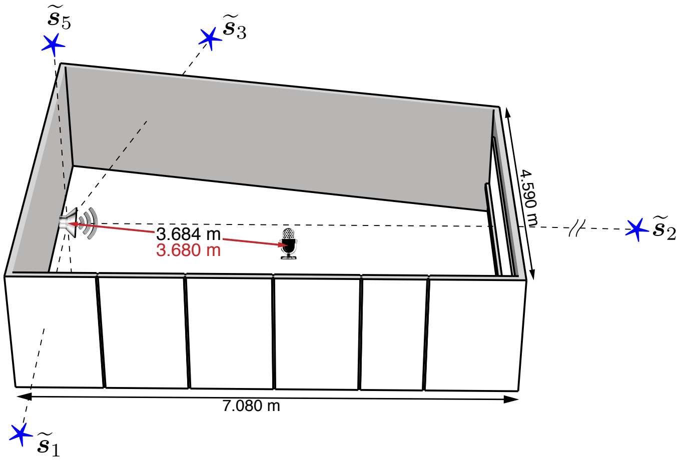


Figure 4.8: Sketch of a room on EPFL campus where the in-room localization experiment is performed. The image sources of the loudspeaker are shown with stars. The image source of the floor ($\widetilde{s}_4$) is not shown for better visualization. The actual distance of the loudspeaker and the microphone is shown in red while the estimated distance is in black.

the north wall, we do not consider the image source for this wall. The EDM of the real and image sources is given as

$$\boldsymbol{D} \approx \begin{bmatrix} 0.00 & 25.40 & 178.48 & 5.91 & 4.66 & 10.38 \\ 25.40 & 0.00 & 203.90 & 55.40 & 30.07 & 35.77 \\ 178.48 & 203.90 & 0.00 & 172.38 & 183.15 & 188.86 \\ 5.91 & 55.40 & 172.38 & 0.00 & 10.58 & 16.28 \\ 4.66 & 30.07 & 183.15 & 10.58 & 0.00 & 28.94 \\ 10.38 & 35.77 & 188.86 & 16.28 & 28.94 & 0.00 \end{bmatrix}.$$

We augment this matrix with 6-tuples of echoes selected from the microphone's RIR. For each combination we find the value of s-stress($\boldsymbol{D}_{\mathsf{aug}}$). The combination that results in the minimum score is selected as the correct combination and the microphone position is found using the estimated permutation of the echoes. As shown in Figure 4.8 position of the microphone was estimated within an error of the order of 1 cm.

4.3.5 The Non-Convex Case: Inverse Method of Images

One of the reasons indoor localization remains a challenge despite numerous attractive applications (see, *e.g.*, [124]) is the occlusion of the line-of-sight path required by many distance-based methods that are efficient outdoor. We demonstrate a simple method that obviates this problem in rooms [159]. We show how to perform indoor localization in non-convex rooms by using the *reverse method of images*—a sequence of reflections of the localized virtual sources. Furthermore, we propose a technique based on gradient descent to refine the position estimate. The refining technique can be used as a building block of a tracking algorithm. The results are valid in both 2D and 3D. While we choose to do illustrations and numerical experiments in 2D purely for simplicity, we note that there do exist “almost-2D” ultrasonic transducers [203] that could be used to obtain echoes from vertical walls only.

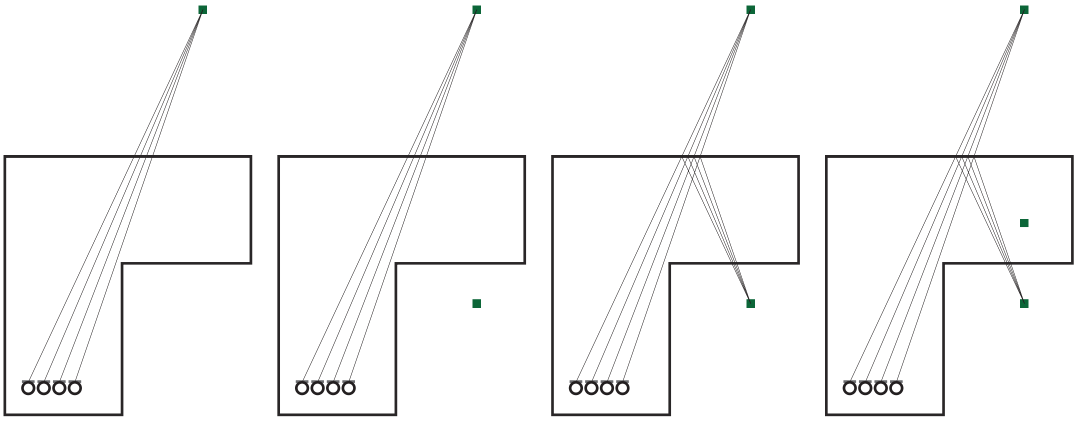


Figure 4.9: Reflecting a localized second order virtual source into the room using the image source method in reverse order.

Similarly to the single-channel localization case, the key here is to treat virtual and real sources equally, thereby transforming the indoor localization problem into a problem of localizing multiple sources in free field using a microphone array. We know from Chapter 3 that the main challenge when performing multilateration from reverberant recordings is to group the echoes that correspond to a single virtual source. This is especially the case with non-compact microphone arrays. To address the labeling problem, we can use the echo sorting algorithm (*cf.* Chapter 3, Algorithm 3.3) or the variant with multilateration described in the previous section.

4.3.6 Reflecting Localized Sources

After finding the location of the virtual source, we can use the knowledge of the room geometry to reflect it back into the room following the method of images in reverse order. After a sequence of reflections, we find the position of the source that generated the localized virtual source.

We explain the reflecting procedure with reference to Figure 4.9. A line is drawn between the virtual source and each of the microphones. Then, we reflect the virtual source across the wall that intersects the drawn lines, and we store the intersection points. If the reflected source is inside the room, we are done. Otherwise, we draw a new set of lines between the previous set of intersection points and the new virtual source, and we reflect the virtual source across the wall that intersects the newly drawn lines. The algorithm is iterated until the reflected source is finally inside the room. In Figure 4.9, the procedure is illustrated for a second order virtual source.

A problem that may occur while applying the inverse method of images is that the lines connecting the virtual source with the microphones intersect more than one wall because of the errors in virtual source estimation. In such case, we may either discard the problematic virtual source, or we can reflect across the wall with the highest number of intersections.

4.3.7 Estimating the Source Position

Thus far, we have localized multiple virtual sources and reflected them back into the room. There are different ways to combine multiple reflected sources into a single position estimate.

For example, we could take the mean location of all sources reflected back into the room, or we could take the one that achieves the best value of the multilateration cost function. However, as the measurement jitter increases, it may happen that wrong echo combinations that do not correspond to any source pass the echo sorting stage. To address this, we propose a scoring scheme suitable for localization in strong measurement jitter.

If the position estimate is close to the source, the simulated RIR from the estimated position is *close* to the recorded RIR. To use this idea to estimate the source position, we define a metric that measures the distance between the two RIRs. For every echo recorded by the microphones, we find the echo closest in time in the simulated RIR, and compute the (squared) 2-norm of the time differences. More precisely, we define the cost function between the RIRs as

$$g(\boldsymbol{x}) = \sum_{m=1}^{M} \sum_{k=1}^{K} e_{ij}^2(\boldsymbol{x}), \tag{4.24}$$

where

$$e_{mk}(\boldsymbol{x}) = \min_{\ell} |t_{mk} - t_{m\ell}^{\star}(\boldsymbol{x})|,$$

time t_{mk} is the delay of kth echo recorded by the mth microphone, $t_{m\ell}^{\star}(\boldsymbol{x})$ is the ℓth echo delay in the simulated RIR for the mth microphone if the source position is $\boldsymbol{x}$, and K is the number of echoes heard by the ith microphone. We can then pick the reflected version of the virtual source that minimizes $g(\boldsymbol{x})$,

$$\boldsymbol{s}_{\text{best}} = \arg\min_{s \in \mathcal{S}} g(\boldsymbol{s}), \tag{4.25}$$

where by $\mathcal{S}$ we denoted the set of all candidates for the source location estimate obtained by reflecting the virtual source back into the room.

4.3.8 Optimizing the Position Estimate

Note that in (4.24) we restricted the search space for the minimizer of g to the discrete set of reflected virtual sources. It is possible to further improve the localization by perturbing the obtained location estimate $\boldsymbol{s}_{\text{best}}$ so that the multilateration cost function in (4.19) is minimized, but this time over the estimated source location and its image sources. Concretely, we want to minimize the following cost function

$$h(\boldsymbol{s}) = \sum_{m=1}^{M} \sum_{k=1}^{K} \left[d_{mk} - \|I_{mk}(\boldsymbol{s}) - \boldsymbol{r}_m\|\right]^2, \tag{4.26}$$

where inner summation goes over the echoes in the recorded responses that we want to match. By $I_{mk}(\boldsymbol{s})$ we denote an image source of $\boldsymbol{s}$ that generates the echo closest to d_{mk} at microphone m. To this image source we associate the sequence of walls that generates it, denoted by $W_{mk}(\boldsymbol{s})$.

One possibility to minimize h is simply by gradient descent. Since we have a very good initial guess for the source location, we expect this to lead to the global optimum with high probability despite the problem not being convex. The gradient can be computed as

$$\nabla h(\boldsymbol{s}) = 2 \sum_{m=1}^{M} \sum_{k=1}^{K} \left[\prod_{w \in W_{mk}(\boldsymbol{s})} (\boldsymbol{I}_2 - 2\boldsymbol{N}_w) \right] \left(\|I_{mk}(\boldsymbol{s}) - \boldsymbol{r}_i\| - d_{mk}\right) \frac{I_{mk}(\boldsymbol{s}) - \boldsymbol{r}_m}{\|I_{mk}(\boldsymbol{s}) - \boldsymbol{r}_m\|},$$

Algorithm 4.1 Localization by the Inverse Method of Images

```
Input: Echo arrival times at the $M$ microphones $T_M$, microphone positions, room geometry
    (normals $\boldsymbol{n}_i$, points $\boldsymbol{p}_i$)
Output: Source location estimate $\widehat{\boldsymbol{s}}$
Localize the set of real and virtual sources $\{\boldsymbol{s}_k\}_{k=1}^K$ using echo sorting (Algorithm 3.3),
for every $k \in \{1, \dots, K\}$ do
    $\widehat{\boldsymbol{s}}_k \leftarrow \boldsymbol{s}_k$
    while $\widehat{\boldsymbol{s}}_k$ not in the room do
        Draw the line segments connecting $\boldsymbol{r}_m$ and $\boldsymbol{s}_k$; they intersect wall $i$
        Mirror the current estimate, $\widehat{\boldsymbol{s}}_k \leftarrow \widehat{\boldsymbol{s}}_k - 2\langle \boldsymbol{p}_i - \widehat{\boldsymbol{s}}_k, \boldsymbol{n}_i \rangle \boldsymbol{n}_i$
    end while
end for
Define the set of mirrored image sources, $\mathcal{S} \leftarrow \{\widehat{\boldsymbol{s}}_1, \dots, \widehat{\boldsymbol{s}}_K\}$,
$\boldsymbol{s}_{\text{best}} = \arg\min_{\boldsymbol{s} \in \mathcal{S}} g(\boldsymbol{s})$                      ▷ $g$ as in (4.24)
$\widehat{\boldsymbol{s}} \leftarrow \arg\min_{\boldsymbol{s} \in \mathbb{R}^3} h(\boldsymbol{s})$          ▷ Gradient descent on (4.26) starting at $\boldsymbol{s}_{\text{best}}$
```

where $\boldsymbol{N}_w \overset{\text{def}}{=} \boldsymbol{n}_w \boldsymbol{n}_w^\top$ and $\boldsymbol{n}_w$ is the normal corresponding to wall w, and we used the expression for the position of an image source (2.62).

Minimizing h can again be motivated by the maximum likelihood argument in Proposition 4.1; however, because we do not have the actual distances between the virtual sources and the microphones, but we rather get them by finding the closest matching echoes in the recorded signals, this minimization should be seen as a heuristic. The complete localization algorithm is summarized in Algorithm 4.1.

4.3.9 Tracking

Source tracking can be performed by repeated localization; the source can be localized independently at each time instant using Algorithm 4.1. However, because the current position of the source depends on previous locations, we can leverage previous estimates to improve the performance.

We propose the following simple method: For the initial position there are no prior estimates, so we localize the source using the algorithm described in Section 3.4. This means that we do 1) echo sorting, 2) virtual source localization, 3) virtual source reflection, and 4) minimization of g. For the remaining time instances, we assume that the source position did not change significantly (by choosing the time interval appropriately), and we localize the source only by solving (4.26) by the gradient descent initialized at the previous position estimate.

4.3.10 Numerical Simulations for Non-Convex Localization

We test the localization algorithm in an L-shaped room shown in Figure 4.10A. The coordinates of the source are $(6 \text{ m}, 7 \text{ m})$, and we use four microphones positioned uniformly at random over the square with corners at $(1 \text{ m}, 1 \text{ m})$, $(1 \text{ m}, 3 \text{ m})$, $(3 \text{ m}, 3 \text{ m})$ and $(1 \text{ m}, 3 \text{ m})$. We stop the simulation after the third order echoes. Note that there is no line-of-sight path between the source and any microphone.

Figure 4.10A shows an outcome of the localization from jitter-free measurements. It can be seen that s-stress prefers positions close to the real source, and that g chooses the best position

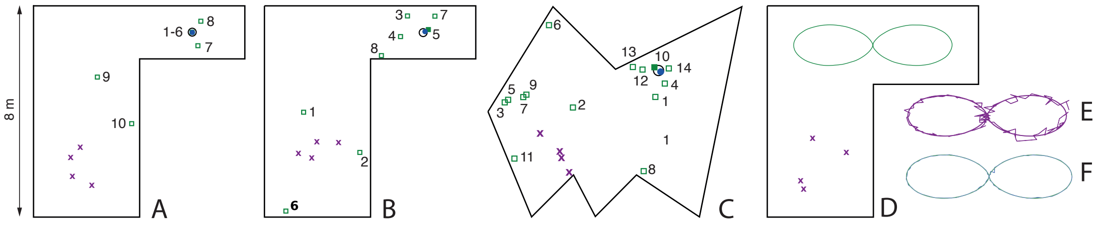


Figure 4.10: Numerical simulations for non-convex localization. In all figures, the purple "x" denotes the microphone position and the black circle depicts the source location. Green squares are the reflected virtual sources ordered by the echo sorting score (s-stress). Smaller indices indicate better scores. The solid square depicts the reflected virtual source with the best g, and the blue dot is obtained by solving (4.26) by a gradient descent initialized at the position of the solid square. (A) Jitter free localization; (B) Measurement jitter from $\mathcal{N}(0, 0.05^2)$; (C) Measurement jitter from $\mathcal{N}(0, 0.1^2)$; (D) Real trajectory in the tracking experiment with jitter with $\sigma = 0.05$; (E) Tracking only by the inverse image method; (F) Tracking by minimizing (4.26).

among the reflected virtual sources. The resulting solid green square overlaps with the blue dot, as in this case both the minimization of g and of h give perfect localization.

Figure 4.10B shows localization with the measurement jitter drawn iid from a centered Gaussian with $\sigma = 0.05$. Although there are reflected sources in the vicinity of the true source position, the ones giving the best s-stress are farther away. Nevertheless, g successfully discriminated the *correct* reflected source (closest to the true position). We observe that minimizing h in (4.26) further improves the position. We tested the algorithm in a more complex room, as shown in Figure 4.10C, with $\sigma = 0.1$. The reflected virtual sources still concentrate around the real source position. Although the positions with the highest localization scores are scattered around the room, g selects the one that is closest to the source. Again, solving minimizing h in (4.26) improves the estimate.

Figure 4.10E and Figure 4.10F show the result of tracking a source that was moving along the curve $[s_1(t)\ s_2(t)]^\top : [0, \infty) \to \mathbb{R}^2$ with $s_1(t) = 4 + 3\cos^3(\pi t/60)$ m and $s_2(t) = 6.5 + 2\sin^3(\pi t/60)$ m, with the jitter variance of $\sigma = 0.05$ m. Figure 4.10E was obtained by going through all of the steps of the localization algorithm at each time instance (echo sorting, multilateration, image source reflecting), but without performing the optimization (4.26). In Figure 4.10F we used *only* the heuristic h. It can be seen that the second, simpler approach performs significantly better.

4.4 Position Calibration of Microphone Arrays

Applications of microphone arrays typically require us to know where the microphones are. If the geometry of the array is not known, we must somehow learn it.

The position calibration problem is particularly important in two groups of applications. The first one is *ad hoc* microphone arrays, for example microphones deployed to run an experiment or to make a recording, or microphone-equipped devices that share the room, such as smartphones, tablets, laptops, or glasses. Another relevant group of applications is in very large, fixed microphone arrays, where precise hand measurements of the microphone positions become impossible. By very large we think of at least several tens, or even hundreds or thousands of microphones [221].

In both groups of applications, measuring microphone positions by hand is slow, imprecise, and impractical. It is also inconvenient to use specialized calibration rigs, *e.g.*, loudspeakers mounted on a fixed construction [182].

Recent practical methods for microphone localization use arbitrarily positioned *acoustic events* for calibration [41, 77, 170]. Raykar and Duraiswami [174] formulate a non-linear least squares problem using at least five loudspeakers, and derive a closed form solution in the case when one loudspeaker is close to a microphone. Crocco, Del Bue and Murino [43] demonstrate an approach that uses low-rank matrix factorization. They, too, derive a closed-form expression for microphone positions for a collocated source and microphone. With sources appearing at known times, the problem is equivalent to multi-dimensional unfolding, with the solution similar to the one in [186].

We focus on calibration from time-of-flight measurements between the microphones and the acoustic events, but we should mention that there are methods that directly measure the pairwise distances between the microphones, typically using the coherence of the diffuse noise [197]. Beyond using distances, some very recent calibration methods exploit efficient representations of reverberant sound fields in terms of, for example, Fourier-Bessel functions [158]. These methods have potential to enable calibration with narrowband sources.

Some methods can also cope with unknown sound emission times. Thrun [202] transposed the Tomasi-Kanade factorization used in computer vision [205] to the microphone position calibration problem; he assumes that the sources are in the far field. Pollefeys and Nister [170] exhibit a similar method that requires no such assumption. Gaubitch, Kleijn and Heusdens [77] additionally allow for unknown internal delays of the microphone processing chain.

On the other hand, these methods require a full set of distance measurements. Furthermore, it is not straightforward to add *a priori* information about the relative geometry of the microphones or of the acoustic events. Such *a priori* information may take any of the following forms:

- Distance between microphones 5 and 10 is 10 cm,
- Calibration events 3, 4, and 5 are all within a ball of 20 cm,
- Distance between microphones 1 and 2 is between 20 and 30 cm,
- Distance between microphones 1 and 2 is larger than the distance between microphones 1 and 3.

Intuition suggests that prior knowledge should improve calibration performance. As we show in numerical experiments, this intuition is confirmed in practice.

To address potentially missing distances between sources and microphones, and to incorporate the prior information about the geometry, we propose to frame microphone localization as a special case of Euclidean distance matrix (EDM) completion.

We do it in two steps. First, we recognize the calibration problem as an instance of multidimensional unfolding (MDU)—a set of tools used for data visualization in psychometrics. The MDU was addressed in detail by Schönemann in 1970 [186].

Second, we recognize that MDU is a special case of the EDM completion problem (Problem 2.1), with a structure complementary to the more thoroughly investigated sensor network localization problem [191, 17]. We solve the EDM completion problem using the methods introduced in Chapter 2. A constraint of this approach is that to use EDMs we need to work with distances, as opposed to distance differences. Thus, we need to assume synchronization between the microphones and the sources.

4.5 Microphone Calibration and Multidimensional Unfolding

We define the microphone localization problem as the task of finding the locations of m microphones given their distances to k acoustic events.

Problem 4.2

> *Denote by* $\boldsymbol{R} = [\boldsymbol{r}_1, \ldots, \boldsymbol{r}_M] \in \mathbb{R}^{d\times M}$ *the unknown microphone locations, and by* $\boldsymbol{S} = [\boldsymbol{s}_1, \ldots, \boldsymbol{s}_K] \in \mathbb{R}^{d\times K}$ *the unknown source locations, where d denotes the ambient dimensionality (usually $d = 2$ or $d = 3$). The distance between the mth microphone and kth source is*
>
> $$\delta_{mk} = \|\boldsymbol{r}_m - \boldsymbol{s}_k\|^2, \tag{4.27}$$
>
> *where $\|\cdot\|$ denotes the Euclidean norm. We collect these distances in the matrix $\boldsymbol{\Delta}$, and the task is to recover $\boldsymbol{R}$ (and $\boldsymbol{S}$) from $\boldsymbol{\Delta}$.*

This standard scenario is described for example in [43]. We first recognize this problem as an instance of metric multidimensional unfolding (MDU) [186]. Indeed, MDU is defined as the problem of localizing a set of points partitioned into two subsets, where we can measure the distances between the points belonging to different subsets, but not between the points in the same subset.

One of the early approaches to the metric MDU is described by Schönemann [186]. We go through the steps of the algorithm, and then explain how we can solve the problem using EDMs. The goal is to make a comparison between the MDU-specific approach and a more general EDM formalization, and to emphasize the universality and simplicity of the EDM approach.

We can write the definition of δ_{mk} (4.27) in matrix form as

$$\boldsymbol{\Delta} = \mathrm{EDM}(\boldsymbol{R}, \boldsymbol{S}) \overset{\mathrm{def}}{=} \mathrm{diag}(\boldsymbol{R}^\top \boldsymbol{R})\boldsymbol{1}^\top - 2\boldsymbol{R}^\top \boldsymbol{S} + \boldsymbol{1}\,\mathrm{diag}(\boldsymbol{S}^\top \boldsymbol{S}). \tag{4.28}$$

Consider now two geometric centering matrices of sizes M and K, denoted $\boldsymbol{J}_M$ and $\boldsymbol{J}_K$. Recall that the geometric centering matrix of size N is defined as

$$\boldsymbol{J}_N = \boldsymbol{I} - \frac{1}{N}\boldsymbol{1}\boldsymbol{1}^\top. \tag{4.29}$$

Using this definition, we have

$$\boldsymbol{R}\boldsymbol{J}_M = \boldsymbol{R} - \boldsymbol{r}_c\boldsymbol{1}^\top,\ \boldsymbol{S}\boldsymbol{J}_K = \boldsymbol{S} - \boldsymbol{s}_c\boldsymbol{1}^\top, \tag{4.30}$$

where $\boldsymbol{r}_c$ is the geometric center of the microphones in $\boldsymbol{R}$, and $\boldsymbol{s}_c$ is the geometric center of the sources in $\boldsymbol{S}$. This means that

$$\boldsymbol{J}_M \boldsymbol{\Delta} \boldsymbol{J}_K = \widetilde{\boldsymbol{R}}^\top \widetilde{\boldsymbol{S}} \overset{\mathrm{def}}{=} \widetilde{\boldsymbol{G}} \tag{4.31}$$

is a matrix of inner products between vectors $\widetilde{\boldsymbol{r}}_M$ and $\widetilde{\boldsymbol{s}}_K$. We used tildes to differentiate this from *real* inner products betwen $\boldsymbol{r}_M$ and $\boldsymbol{s}_K$, because in (4.31), the points in $\widetilde{\boldsymbol{R}}$ and $\widetilde{\boldsymbol{S}}$ are referenced to different coordinate systems. The centroids $\boldsymbol{r}_c$ and $\boldsymbol{s}_c$ generally do not coincide. There are different ways to decompose $\widetilde{\boldsymbol{G}}$ into a product of two full rank matrices $\boldsymbol{A}$ and $\boldsymbol{B}$,

$$\widetilde{\boldsymbol{G}} = \boldsymbol{A}^\top \boldsymbol{B}. \tag{4.32}$$

We could for example use the SVD, $\widetilde{\boldsymbol{G}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, and set $\boldsymbol{A}^\top = \boldsymbol{U}$ and $\boldsymbol{B} = \boldsymbol{\Sigma}\boldsymbol{V}^\top$. Any two such decompositions are linked by some invertible transformation $\boldsymbol{T} \in \mathbb{R}^{d\times d}$,

$$\widetilde{\boldsymbol{G}} = \boldsymbol{A}^\top \boldsymbol{B} = \widetilde{\boldsymbol{R}}^\top \boldsymbol{T}^{-1}\boldsymbol{T}\widetilde{\boldsymbol{S}}. \tag{4.33}$$

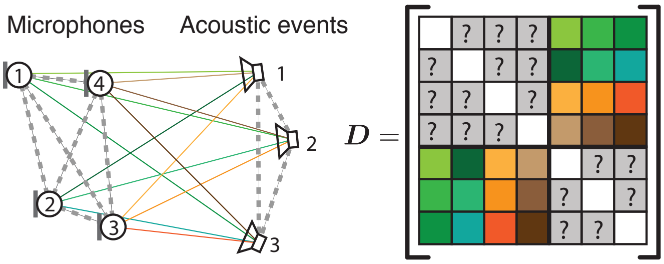


Figure 4.11: Microphone calibration as an example of MDU. We can measure only the propagation times from acoustic sources at unknown locations, to microphones at unknown locations. The corresponding revealed part of the EDM has a particular off-diagonal structure, leading to a special case of EDM completion.

We can now write down the conversion rule between the sought microphone positions and the matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ which we can compute from the distance measurements $\boldsymbol{\Delta}$:

$$\begin{aligned} \boldsymbol{R} &= \boldsymbol{T}^\top \boldsymbol{A} + \boldsymbol{r}_c \boldsymbol{1}^\top \\ \boldsymbol{S} &= (\boldsymbol{T}^{-1})^\top \boldsymbol{B} + \boldsymbol{s}_c \boldsymbol{1}^\top, \end{aligned} \tag{4.34}$$

where $\boldsymbol{A}$ and $\boldsymbol{B}$ are computed according to (4.32). Because we cannot reconstruct the absolute position of the point configuration, we can arbitrarily set $\boldsymbol{r}_c = 0$, and $\boldsymbol{s}_c = \alpha \boldsymbol{e}_1$. Recapitulating, we have that

$$\boldsymbol{\Delta} = \operatorname{EDM}\left(\boldsymbol{T}^\top \boldsymbol{A},\ (\boldsymbol{T}^{-1})^\top \boldsymbol{B} + \alpha \boldsymbol{e}_1 \boldsymbol{1}^\top\right), \tag{4.35}$$

and the problem is reduced to computing $\boldsymbol{T}$ and α so that (4.35) hold, or in other words, so that the right hand side be consistent with the data $\boldsymbol{\Delta}$. We reduced MDU to a relatively small problem: In 3D, we need to compute only ten scalars. Schönemann [186] gives an algebraic method to find these parameters, and mentions the possibility of least squares, while Crocco, Del Bue and Murino [43] clearly outline an approach using non-linear least squares. We should add here that their method differs from Schönemann's in yet another way: Instead of subtracting the centers of $\boldsymbol{R}$ and $\boldsymbol{S}$ using the geometric centering matrix, they subtract the first point. As discussed in Chapter 2, this can be modeled as a multiplication of the point set by a matrix of the form

$$\boldsymbol{I} - \boldsymbol{e}_1 \boldsymbol{1}^\top. \tag{4.36}$$

In both cases, the described procedure seems quite convoluted. Rather, we see MDU as a special case of matrix completion.

4.5.1 Unfolding as Matrix Completion

We now frame Problem 4.2 (MDU) as a special case of Problem 2.1 (EDM completion). Let the microphones and the sources be represented by a set of $N = K + M$ points, ascribed to the columns of matrix $\boldsymbol{X} = [\boldsymbol{R}\ \boldsymbol{S}]$. Then $\operatorname{EDM}(\boldsymbol{X}^\top \boldsymbol{X})$ has a special structure as illustrated in Figure 4.11,

$$\operatorname{EDM}(\boldsymbol{X}) = \begin{bmatrix} \operatorname{EDM}(\boldsymbol{R}) & \operatorname{EDM}(\boldsymbol{R}, \boldsymbol{S}) \\ \operatorname{EDM}(\boldsymbol{S}, \boldsymbol{R}) & \operatorname{EDM}(\boldsymbol{S}) \end{bmatrix}, \tag{4.37}$$

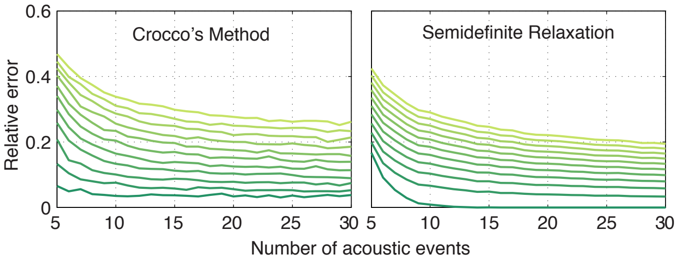


Figure 4.12: Comparison of two algorithms applied to multidimensional unfolding with varying number of acoustic events K and noisy distances. For every number of acoustic events, we generated 1000 realizations of $M = 20$ microphone locations uniformly at random in a unit cube. In addition to the number of acoustic events, we varied the amount of random jitter added to the distances. Jitter was drawn from a centered uniform distribution, with the level increasing in the direction of the arrow, from $\mathcal{U}[0, 0]$ (no jitter) for the darkest curve at the bottom, to $\mathcal{U}[-0.15, 0.15]$ for the lightest curve at the top, in 11 increments. For every jitter level, we plotted the mean relative error $\|\widehat{\boldsymbol{D}} - \boldsymbol{D}\|_F / \|\boldsymbol{D}\|_F$ for all algorithms.

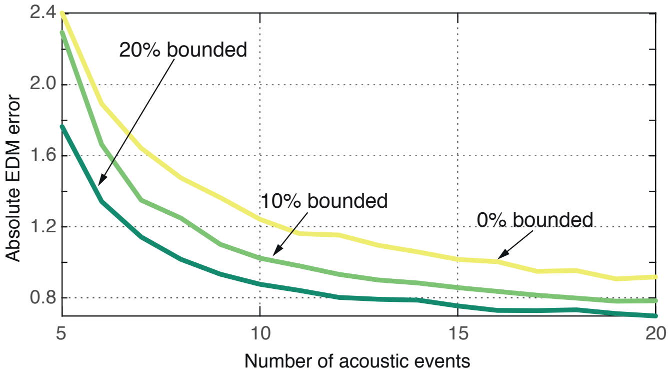


Figure 4.13: Comparison of the influence of prior information on the reconstruction accuracy. For every number of acoustic events, we generated 500 realizations of 15 microphones inside the unit cube (acoustic event were also generated inside the unit cube). Jitter of ± 7 cm was added to the distances measurements and then varying percentage of microphone pairwise distances was revealed within ± 15 % of the true distance. The curves show the absolute reconstruction error $\|\widehat{\boldsymbol{D}} - \boldsymbol{D}\|_F$. Mean value of the Frobenius norm of the true EDM was around 13 for 20 acoustic events, so the absolute error of 1 can be considered as successful localization.

where the upper-right and the lower-left blocks are measured, and the diagonal blocks are unknown. Thus we define the mask matrix for MDU as

$$\boldsymbol{W}_{\mathrm{MDU}} \overset{\mathrm{def}}{=} \begin{bmatrix} \boldsymbol{0}_{M \times M} & \boldsymbol{1}_{M \times K} \\ \boldsymbol{1}_{K \times M} & \boldsymbol{0}_{K \times K} \end{bmatrix}. \tag{4.38}$$

With this matrix, we can simply invoke the SDR in Algorithm 2.5 (Chapter 2).

4.5.2 Integrating Prior Knowledge

Recall the semidefinite relaxation with trace maximization for completing noisy EDMs (4.39),

$$\begin{array}{ll} \underset{\boldsymbol{H}}{\text{maximize}} & \operatorname{trace}(\boldsymbol{H}) - \lambda \| \boldsymbol{W} \circ (\widetilde{\boldsymbol{D}} - \mathcal{K}(\boldsymbol{V}\boldsymbol{H}\boldsymbol{V}^\top)) \|_F, \\ \text{subject to} & \boldsymbol{H} \in \mathbb{S}_+^{N-1}. \end{array} \tag{4.39}$$

The attractiveness of this formulation with the Gram matrix, as opposed to explicitly writing out the stress functions, is in its flexibility: it is easy to insert many useful constraints on the distances. Furthermore, unlike with the methods that are developed specifically for microphone calibration, if we miss one or more source-microphone distances, (4.39) still works with a trivial modification of the mask matrix $\boldsymbol{W}$. Such situations arise commonly in practice.

It is easy to insert estimates or bounds on the distances. For example, we may have a rough idea of the distances between some pairs of microphones, as well as a rough idea of the distances between some pairs of acoustic events. All of these constraints are simply linear constraints on $\boldsymbol{D} = \mathcal{K}(\boldsymbol{V}\boldsymbol{H}\boldsymbol{V}^\top)$. Let us give some examples. Upper and lower bounds can be specified as follows

$$\begin{aligned} \boldsymbol{W}_L \circ [\mathcal{K}(\boldsymbol{V}\boldsymbol{H}\boldsymbol{V}^\top) - \boldsymbol{L}] \geqslant 0 \\ \boldsymbol{W}_U \circ [\mathcal{K}(\boldsymbol{V}\boldsymbol{H}\boldsymbol{V}^\top) - \boldsymbol{U}] \leqslant 0, \end{aligned} \tag{4.40}$$

where matrices $\boldsymbol{W}_L$ and $\boldsymbol{W}_U$ select the elements for which we have bounds, and $\boldsymbol{L}$ and $\boldsymbol{U}$ contain the bounds.

We can also plug in ordinal information about the distances. For example, we can require that the distance between the microphones i and j be smaller than the distance between the microphones i and k. This is written simply as

$$\boldsymbol{e}_i^\top \mathcal{K}(\boldsymbol{V}\boldsymbol{H}\boldsymbol{V}^\top)(\boldsymbol{e}_j - \boldsymbol{e}_k) \leqslant 0, \tag{4.41}$$

where $\boldsymbol{e}_i$ is the ith canonical basis vector. Many other useful constraints can be modeled as linear constraints on $\mathcal{K}$.

4.5.3 Numerical Experiments

We demonstrate the usefulness of EDMs for the microphone calibration problem in two different scenarios. First, we address the problem of localizing the microphone array from the set of pairwise distance measurements in the presence of noise, and compare the SDR with the method of Crocco, Del Bue and Murino [43]. The results are illustrated in Figure 4.12. We observe that for the noiseless case, the SDR performs better, giving perfect reconstruction for all but the lowest number of acoustic events. It also yields more accurate reconstructions at higher noise levels. When the number of acoustic events is very low, the method from [43] performs better; SDR retrieves solutions whose embedding dimension is too high.

Second, we explore the effect of adding prior information on the distances between the microphones in Figure 4.13. We add considerable jitter to measurements (±7 cm within a meter is surpassed by any method for measuring distance using times of flight). We then add upper and lower bounds at 15% of the true distance for varying percentage of the pairwise microphone distances. It is clearly observed that adding prior information can improve the localization performance. A comparison with other approaches is not possible as they do not allow to easily integrate such constraints.

4.6 How to Localize Ten Microphones in a Finger Snap (With a Little Help From Echoes)

We have seen in the previous section that a compelling method to calibrate the positions of microphones in an array is with sources at unknown locations; remarkably, it is possible to reconstruct the locations of both the sources and the receivers, if their number is larger than some prescribed minimum. All existing methods, based on times of arrival or time differences of arrival, only exploit the direct paths between the sources and the receivers. In the existing approaches, the room reverberation is either ignored or considered detrimental.

In this section we present a proof-of-concept of using echoes to considerably reduce the number of sources needed for calibration. We demonstrate that being in a room reduces the number of acoustic events needed for calibration, even if we do not know how the room looks like or how the microphone array is positioned and oriented inside the room. It is somewhat surprising to note that the echoes help in the calibration despite not knowing where they are coming from. Supposing that the positions of the walls are unknown, the location of the source is unknown, and the locations of all the microphones are unknown, we are still able to estimate all these parameters. In fact, we can do the estimation from a single finger snap.

Our algorithm outputs two particularly useful *by-products*: 1) information about the room shape (as we also localize virtual sources), and 2) the array's absolute position in the room, not available with other calibration methods. The proposed procedure is in a way a *total calibration*—we learn about microphones, sources and reflectors without knowing them *a priori*. We show through numerical simulations that the algorithm can indeed localize ten microphones with a single sound source.

4.6.1 Anechoic Calibration

As a building block in our approach, we use an algorithm for anechoic calibration. Any of the algorithms mentioned in the previous section will do. Assume that the sources produce some impulsive sound that the microphones record, and whose time-of-arrival (TOA) we can estimate (up to a possibly unknown offset); assume further that the microphones are synchronized.

We denote the source positions by $\{\boldsymbol{s}_k\}_{k=1}^K$, and the microphone positions by $\{\boldsymbol{r}_m\}_{m=1}^M$. An offset time τ_k is associated with the kth source. Then we can express the measurements as

$$\vartheta_{km} = c\,\tau_k + \|\boldsymbol{s}_k - \boldsymbol{r}_m\|_2. \tag{4.42}$$

We collect the measurements in a matrix $\boldsymbol{\Theta} = \left[\vartheta_{km}\right]_{k,m=1}^{K,M}$.

As announced, we assume the existence of a module—a black box as far as we are concerned—that we denote **Calibrate**, and that computes the estimates of the unknown microphone and source locations $\boldsymbol{R} \overset{\text{def}}{=} [\boldsymbol{r}_1,\ \ldots,\ \boldsymbol{r}_M]$, $\boldsymbol{S} \overset{\text{def}}{=} [\boldsymbol{s}_1,\ \ldots,\ \boldsymbol{s}_K]$, and offsets $\boldsymbol{\tau} = [\tau_1,\ \ldots,\ \tau_K]^T$ from $\boldsymbol{\Theta}$. We can write

$$(\widehat{\boldsymbol{R}}, \widehat{\boldsymbol{S}}, \widehat{\boldsymbol{\tau}}, \epsilon) = \textbf{Calibrate}(\boldsymbol{\Theta}), \tag{4.43}$$

where $\epsilon \geqslant 0$ denotes some measure of fit. The measure of fit is computed as the discrepancy between the measured data and the data that would have been generated by sensors at estimated positions,

$$\epsilon = \sum_{k=1}^{K} \sum_{m=1}^{M} \left| \vartheta_{km} - (\|\widehat{\boldsymbol{s}}_k - \widehat{\boldsymbol{r}}_m\|_2 + c\,\widehat{\tau}_k) \right|^2 \tag{4.44}$$

If $\widehat{\boldsymbol{R}}$, $\widehat{\boldsymbol{S}}$ and $\widehat{\boldsymbol{\tau}}$ perfectly generate $\boldsymbol{\Theta}$, then $\epsilon = 0$.

Any algorithm behind the **Calibrate** component is associated with a certain minimal number of microphones and sources required for estimation, call them $M_{\min}$ and $K_{\min}$. Typically, $K_{\min}$ is a (non-increasing) function of $M_{\min}$. A particular consequence of this is that the minimal number of columns of $\boldsymbol{\Theta}$ is $M_{\min}$, and the smallest corresponding number of rows $K_{\min}$. When the source offsets are known or all equal, we can swap $M_{\min}$ and $K_{\min}$ by invoking the duality.

4.6.2 Zero-Knowledge Indoor Calibration

In a room, or more generally in the presence of acoustic reflectors, the sources $\{\boldsymbol{s}_k\}$ generate reflections, and the reflections are equivalent to virtual sources (mirror images of the real sources across corresponding walls).

But this means that by virtue of echoes, we get additional sources *for free*. Normally, we would consider these echoes to be a nuisance because, unlike in the single-channel localization case from Section 4.3.2, we do not know where the virtual sources are located (remember that the room shape is unknown). But in the microphone position calibration problem (Problem 4.2), we do not know the locations of the real sources either. Thus, virtual sources at unknown locations are just as good (or just as bad) as real sources at unknown locations.

A challenge that again appears in this setting is that we cannot address each virtual source individually—they are not labeled, and with multiple walls they are heard by microphones in different orders. This problem does not appear with real calibration events, as they are well separated in time. Thus, we need to label the echoes by performing *echo sorting*, introduced in Chapter 3. There, the need to disambiguate echoes (virtual sources) arises when estimating the shape of a room from sound. However, the problems are not the same—in the scenario therein, we assume that we know the relative geometry of the microphone array; in the zero-knowledge calibration problem, we do not know it. This means that the minimal number of microphones and the minimal number of sources will be higher, as reflected by $K_{\min}$ and $M_{\min}$. Moreover, we cannot perform the assignment echo-by-echo, the way we did in Algorithm 3.3; rather, we must assign all echoes at once. This results in a significantly higher computational cost.

The principle at play is similar to the one used in Chapter 3: Provided we have enough noiseless measurements ϑ_{km}, the equations (4.42) yield a unique solution for locations and offsets. That is, these are the only $\widehat{\boldsymbol{R}}$, $\widehat{\boldsymbol{S}}$ and $\widehat{\boldsymbol{\tau}}$ that could have generated $\boldsymbol{\Theta}$. Depending on the solution strategy (*e.g.*, solving an optimization program), any wrong permutation or assignment of echoes will lead to an unsolvable system (4.42), or will yield a wrong solution that cannot recreate the measurements $\boldsymbol{\Theta}$.

The goal is to find the best fit among all possible echo assignments. This can be achieved by running **Calibrate** for different echo assignments, and taking as the correct assignment the one with the smallest ϵ. The described procedure is summarized in Algorithm 4.6.4.

Performing the combinatorial search is feasible for small array sizes. For large arrays, however, the number of combinations becomes too big, and we need to do something else. In this case, we can bootstrap the method by first running it for one or more sub-arrays of the big array. Depending on the target application, we might even have an idea about groups of microphones that are spatially close (this will be the case for large fixed arrays). Knowing which microphones are close in space is relevant, as proximity makes it more likely that the microphones will have picked up the same echoes. In spatially large arrays, it is not guaranteed that all the microphones will hear all the echoes.

The estimation can be performed one acoustic event at a time (*e.g.* a finger snap). This is useful, as we know that the time offset τ will be the same for all virtual events corresponding to a single real event (and it will be equal to the offset of that event). Structured information like this can be exploited in the design of the **Calibrate** module.

4.6.3 Minimal Infrastructure for Calibration

We can use a degree-of-freedom counting argument as in [77] to determine the smallest number of microphones and sources necessary for the calibration. Every microphone brings in d unknowns for each of its coordinates where d is the ambient dimensionality, while every source brings in d unknowns (coordinates) in the synchronized case, and $d + 1$ unknowns (coordinates and the offset τ) in the asynchronous case. In applications d is typically 2 or 3.

On the other hand, we get a measurement for every source-microphone pair (for every TOA measurement), so that the number of measurements is MK, and we need this number to be larger than the total number of unknowns. Note further that we can fix the location of one microphone and the rotation of the remaining points around this microphone. This takes out a total of $d + \binom{d}{2}$ degrees of freedom, resulting finally in

$$K_{\text{sync}} \geqslant \frac{d(2M-d-1)}{2(M-d)} \quad \text{and} \quad K_{\text{async}} \geqslant \frac{d(2M-d-1)}{2(M-d-1)}. \tag{4.45}$$

The above bounds are fundamental, and for fewer acoustic events the solution set is not zero-dimensional. It is important to note that minimal solvers do not yield a unique solution, but rather a zero-dimensional variety. For example, Kuang *et al.* [115] show that the minimal solver in 3D ($M = 4,\ K = 6$) gives 38 solutions. Some of these will be complex, but even the number of real solutions is greater than one. To complicate things further, the number of real solutions depends on the problem instance. In order to obviate these complications, we consider the minimal number of microphones and sources as a property of the method used to calibrate.

We use this example to show that something remarkable can happen in a room. Suppose that $M = 10$. In this case, we compute that $K \geqslant 4$; that is, we need at least four sources. Now imagine that in a room we have a single acoustic event, and that we can get at least three echoes. Together with the direct sound, we get enough measurements (real and virtual) to actually calibrate the microphone array, and to determine its absolute orientation with respect to the walls. This is true in spite of the fact that we do not know the room, the microphone locations, the source location nor the source timing. In this case we only need to estimate a single emission time, as it will be the same for all the image sources. We note that in this case the solution is unique with high probability.

4.6.4 Algorithm

The algorithm, summarized in Algorithm 4.6.4 simply tries all plausible assignments, and chooses the one with the smallest ϵ as measured by the **Calibrate** module as the estimate. In order to reduce the number of combinations as much as possible, the algorithm is specified for the minimal number of microphones and sources solvable by a particular chosen **Calibrate** module.

The assignment problem in zero-knowledge calibration can be seen as the problem of labeling edges in a complete bipartite graph. This perspective is illustrated in Figure 4.14.

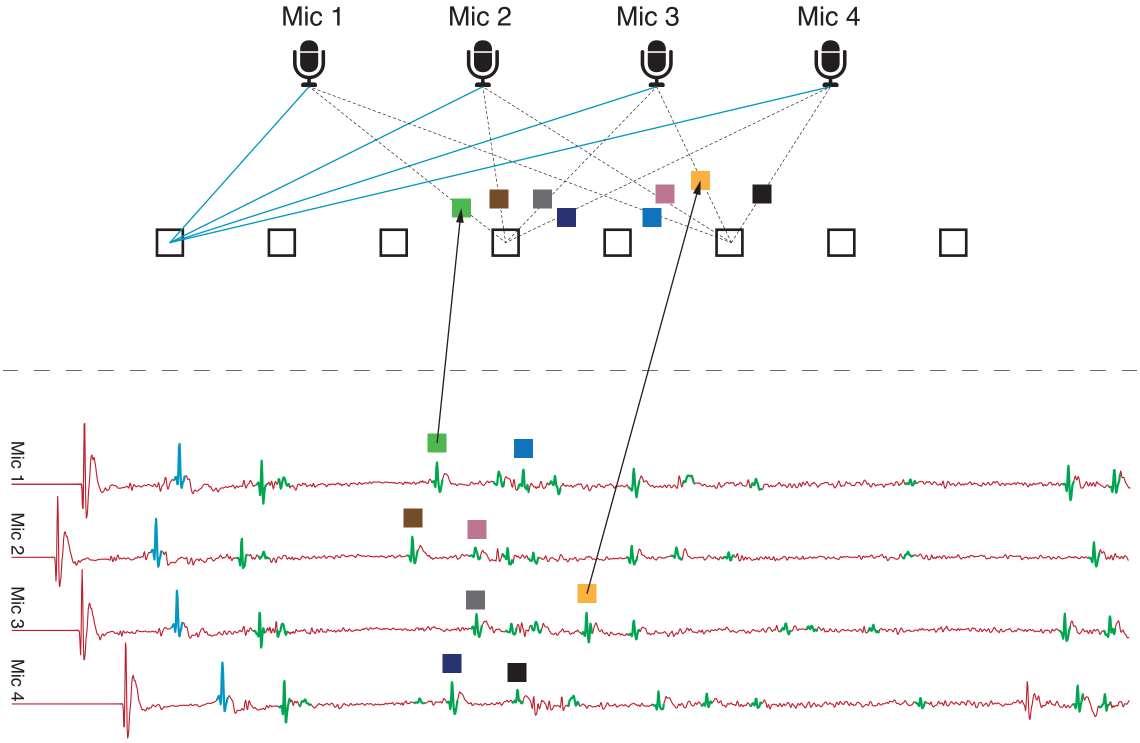


Figure 4.14: Illustration of the zero-knowledge echo sorting procedure.

In the noiseless case or with sufficiently small noise, the **Calibrate** box can simply be a polynomial system solver, and thus very fast for such a small problem. The question then becomes how many times we need to call the solver; this is the subject of the following proposition:

Proposition 4.4

Associate with a particular **Calibrate** *module the minimal number of microphones* M_{min} *and sources* K_{min}. *Suppose further that we detect* K *echoes per microphone. Then the number of assignments to test equals*

$$\left[\frac{(K-1)!}{(K-K_{\text{min}})!}\right]^{M_{\text{min}}-1} \tag{4.46}$$

This number looks daunting, but in small-dimensional cases it is still manageable. For example, in 2D the number of combinations to test for the minimal solver is just $O(K^4)$. For the minimal solver in 3D however, it is $O(K^{15})$ which rapidly becomes impractical, and various heuristics must be used to trim down the search spaces. Some examples of heuristics are:

- Combine the echoes only within a temporal window corresponding to the array size,
- Assume only a small number of echo swaps can occur per microphone,

Algorithm 4.2 Zero-Knowledge Calibration

```
Input: Echo arrival times at the $M$ microphones $\mathcal{T}_m$
Output: Microphone locations $\boldsymbol{r}_m$, real source location $\boldsymbol{s}$, image source locations $\boldsymbol{s}_k$
Fix the origin and the rotation (6 degrees of freedom)
Label the first arriving echo in all microphones as the direct sound
Label the first $K_{\text{min}} - 1$ echoes in $\boldsymbol{r}_1$ arbitrarily and assign them to $\boldsymbol{t}_1$
$\epsilon_{\text{best}} \leftarrow +\infty$
for each assignment tuple $\boldsymbol{t}_2 \in \mathcal{A}_{(K_{\text{min}}-1)}(\mathcal{T}_2), \ \ldots, \ \boldsymbol{t}_{M_{\text{min}}} \in \mathcal{A}_{(K_{\text{min}}-1)}(\mathcal{T}_{M_{\text{min}}})$ do
    Construct the measurement matrix $\boldsymbol{\Theta}$
    Compute the estimate $[\widehat{\boldsymbol{R}}, \widehat{\boldsymbol{S}}, \widehat{\boldsymbol{\tau}}, \epsilon] \leftarrow \textbf{Calibrate}(\boldsymbol{\Theta})$,
    if $\epsilon < \epsilon_{\text{best}}$ then
        $[\widehat{\boldsymbol{R}}_{\text{best}}, \widehat{\boldsymbol{S}}_{\text{best}}, \widehat{\tau}_{\text{best}}, \epsilon_{\text{best}}] \leftarrow [\widehat{\boldsymbol{R}}, \widehat{\boldsymbol{S}}, \widehat{\boldsymbol{\tau}}, \epsilon]$
    end if
end for
Use the localized microphone and echo sorting to localize the remaining microphones and sources
```

- Assume echo swaps occurred at a limited number of microphones,
- Normalize by the decay and discriminate first-order peaks by the magnitude,
- Order the echo assignments by some heuristic *a priori* likelihood and stop as soon as we get a score below a prescribed threshold.

The most potent heuristic is the one about the array size, and that is the one used in practice. To clearly quantify the savings, we state the following proposition:

Proposition 4.5

Assume that the diameter of the microphone array is sufficiently small so that any echo received by a given microphone can be combined with at most J adjacent echoes received by any other microphone. Then the number of combinations to test is at most $J^{(K_{min}-1)(M_{min}-1)}$.

Example 4.1

Consider the minimal solver in 3D, with $M_{min} = 4$ and $K_{min} = 6$, and suppose that we detect 10 echoes per microphone. Using no heuristics, the number of combinations to test is $\approx 3.5 \times 10^{12}$ which is intractable on a desktop computer.

If we further assume that the array is not too large, so that $J = 3$, the number of combinations to test becomes at most 14 million, which can be tested in reasonable time.

4.6.5 Numerical Experiment

We ran numerical simulations using only one acoustic event to localize ten microphones. We simulated a shoebox room with dimensions $W = 5$ m, $L = 6$ m, $H = 3$ m, using the image source model, up to second-order reflections. We experimented with random microphone array geometries and with different numbers of microphones. First six echoes were used in conjunction with the array size heuristic to obtain the reconstructions.

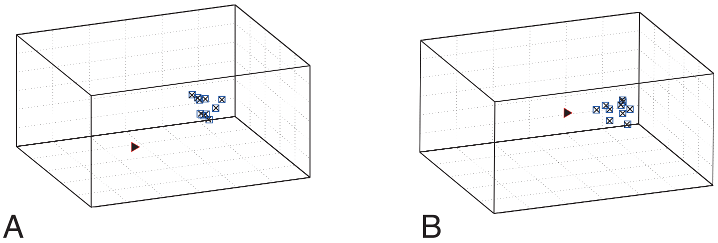


Figure 4.15: (A) and (B) Two typical reconstruction results with $M = 10$ microphones randomly placed inside a box approximately 1m × 1m × 0.5m large. We emphasize that the room dimensions (5m × 6m × 3m) and the room shape is not assumed known. Red-black triangle represents the source location. Small black crosses are true microphone locations, while blue squares denote the estimated locations.

Simulations confirm that it is possible to obtain accurate estimates of microphone positions by using only a single source. The room shape or dimensions are considered unknown by the algorithm. Despite this, we obtain a full reconstruction of the source and microphone locations, as well as their absolute pose inside the room (more precisely, relative to the localized image sources). We also obtain the positions of walls corresponding to these image sources. Two reconstruction examples are shown in Figures 4.15A and 4.15B, for random microphone arrays comprising ten microphones.

4.A Proof of Proposition 4.1

The distance measurements are given as

$$d_m = \|\boldsymbol{r}_m - vs\| + n_m, \tag{4.47}$$

where $n_m \sim \mathcal{N}(0, \sigma^2)$. Since n_m are assumed independent, the joint probability distribution can be written as

$$p(d_1, \ldots, d_m; \boldsymbol{s}) = \prod_{m=1}^{M} \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-(d_m - \|\boldsymbol{r}-\boldsymbol{s}\|)^2/\sigma^2} \tag{4.48}$$

so that the log-likelihood is

$$L(\boldsymbol{s}; d_1, \ldots, d_m) = \sum_{m=1}^{M} -\ln(\sigma\sqrt{2\pi}) - \frac{(d_m - \|\boldsymbol{r}_m - \boldsymbol{s}\|)^2}{\sigma^2}. \tag{4.49}$$

The maximum likelihood estimator is then obtained as

$$\widehat{\boldsymbol{s}} = \arg\max_{\boldsymbol{s}} L(\boldsymbol{s}; d_1, \ldots, d_m) = \arg\min \sum_{m=1}^{M} \big(d_m - \|\boldsymbol{r}_m - \boldsymbol{s}\|\big)^2. \tag{4.50}$$

Chapter 5

Sampling Sparse Signals on the Sphere[*]

> Because the world is round it turns me on,
> Because the world is round...
>
> *The Beatles*

5.1 Introduction

Our initial motivation to look at signal processing on the sphere was to solve source localization problems with spherical sensor geometries; a common example are spherical microphone arrays. This turned into a development of theory and algorithms for sampling sparse signals on the sphere, which we describe in this chapter.

Numerous signals live on a sphere. Take for example any signal defined on Earth's surface [71, 189, 9]. Signals from space measured on Earth [97, 132] also have a spherical domain. In acoustics, spherical microphone arrays output a time-varying signal supported on a sphere [145, 98], while in diffusion weighted magnetic resonance imaging fiber orientations live on a sphere [207]. In practice, we only have access to a finite number of samples of such signals, so it is important to study sampling and reconstruction of spherical signals.

Just as signals in Euclidean domains can be expanded via sines and cosines [214], one can naturally represent spherical signals in the Fourier domain via *spherical harmonics* [65]. A signal is *bandlimited* if it is a linear combination of finitely many spherical harmonics.

In this chapter, we study the problem of sampling localized spikes on the sphere; in the limit, the spikes become Dirac delta functions. Such sparse signals on the sphere are encountered in

---

[*] This chapter is a result of a joint work with Yue M. Lu [53].

many problems. For example, various acoustic sources are well-approximated by point sources; the directional distribution of multiple sources is then a finite collection of spikes. Stars in the sky observed from Earth are angular spikes, and so are plume sources on Earth.

Localized spikes are not bandlimited, so the bandlimited sampling theorems [65, 141, 107] do not apply. In this chapter, we propose an algorithm to perfectly reconstruct collections of spikes from their lowpass-filtered observations. Our algorithm efficiently reconstructs K spikes when the bandwidth of the lowpass filter is at least $K + \sqrt{K}$.

5.1.1 Prior Art

Sampling bandlimited signals on the sphere has been studied extensively: For signals bandlimited to spherical harmonic degree L, Driscoll and Healy [65] proposed a sampling theorem that requires $4L^2$ spherical samples. The best exact general purpose sampling theorem due to McEwen and Wiaux uses $2L^2$ samples [141]. Recently, Khalid, Kennedy and McEwen devised a stable sampling scheme that requires the optimal number of samples, L^2 [107].

Our work is in the same spirit as finite rate-of-innovation (FRI) sampling, introduced by Vetterli, Marziliano, and Blu [215]. They showed that a stream of K Diracs on the line can be efficiently recovered from $2K + 1$ samples. Initially developed for 1D signals, the original FRI sampling was extended to 2D and higher-dimensional signals in [136, 187], and its performance was studied in noisy conditions [137, 63].

In a related work [47, 48], Deslauriers-Gauthier and Marziliano proposed an FRI sampling scheme for signals on the sphere, reconstructing K Diracs from $4K^2$ samples. Their motivating application is the recovery of the fiber orientations in diffusion weighted magnetic resonance imaging [46, 207]. They further show that if only $3K$ spectral bins are active, the required number of samples can be reduced to $3K$. Sampling at this lower rate, however, relies on the assumption that we can apply arbitrary spectral filters to the signal before sampling. This is known as spatial anti-aliasing—a procedure that is generally challenging or impossible to implement in most applications involving spherical signals, where we only have access to finite samples of the underlying continuous signals.[1]

In many applications, the sampling kernels (*i.e.*, the lowpass filters) through which we observe the spikes are provided by some underlying physical process (*e.g.*, point spread functions and Green's functions). These kernels are often approximately bandlimited, but we cannot further control or design their spectral selectivity. This impossibility of arbitrary spatial filtering suggests that our goal is to reduce the required bandwidth, or more practically, to maximize the number of spikes that we can reconstruct at a given bandwidth.

Recently, Bendory, Dekel and Feuer proposed a spherical super-resolution method [14, 15], extending the results of Candès and Fernandez-Granda [30] to the spherical domain. They showed that an ensemble of Diracs on the sphere can be reconstructed from projections onto a set of spherical harmonics by solving a semidefinite program, provided that the Diracs satisfy a minimal separation condition. When the Diracs are constrained to a discrete set of locations, their formulation allows them to bound the recovery error in the presence of noise. Our non-iterative (thus very fast) algorithm based on FRI does not require any separation between the Diracs. We also allow the weights to be complex, which may be important in applications (for an example on sound source localization, see Section 5.4.3). On the other hand, we need to assume

---

[1] This is not to be confused with spatial anti-aliasing in image downsampling, where we do have access to all pixels.

that the number of Diracs is known a priori (or that it can be estimated through other means), whereas in [14, 15] no such assumption is necessary.

5.1.2 Outline and Main Contributions

We start by reviewing some basic notions of harmonic analysis on the sphere in Section 5.2. We then present the main result of this work in Section 5.3 : A collection of K Diracs on the sphere can be reconstructed from its lowpass filtered version, provided that the bandwidth of the sampling kernel is at least $K + \sqrt{K}$. This bandwidth requirement also implies that $(K + \sqrt{K})^2$ spatial samples taken at generic locations suffice to reconstruct the K Diracs. We establish this result by constructing a new algorithm for spherical FRI sampling. Compared to $4K^2$ samples as required in a previous work [47], our algorithm reduces the numbers of samples via a more efficient use of the available spectrum. For large K, the required number of samples is reduced by a factor of up to 4. The proposed algorithm is first developed for the noiseless case. Procedures to improve the robustness of the algorithm in noisy situations are presented in Section 5.3.5, and we compare the performance of the algorithm with the Cramér-Rao lower bound [101] in Section 5.3.6. Section 5.4 presents the applications of the proposed algorithm to three problems: 1) sampling diffusion processes on the sphere, 2) shot noise removal, and 3) sound source localization. These diverse applications demonstrate the usefulness and versatility of our results.

5.2 Harmonic Analysis on the Sphere and Problem Formulation

5.2.1 Spherical Harmonics

We briefly recall the definitions of spherical harmonics and spherical convolution. The 2-sphere is defined as the locus of points in $\mathbb{R}^3$ with unit norm,

$$\mathbb{S}^2 \stackrel{\text{def}}{=} \left\{ \boldsymbol{x} \in \mathbb{R}^3 \mid \boldsymbol{x}^\top \boldsymbol{x} = 1 \right\}.$$

In what follows, we often use ξ to represent a generic point on the sphere. In addition to the standard Euclidean representation $\xi = [x,\ y,\ z]^\top$, points on $\mathbb{S}^2$ can also be conveniently parameterized by angles of colatitude and azimuth, *i.e.*, $\xi = (\theta, \phi)$, with θ measured from the positive z-axis, and ϕ measured in the xy plane from the positive x-axis. The two equivalent representations are related by the following conversion:

$$\begin{aligned} x &= \sin(\theta)\cos(\phi), \\ y &= \sin(\theta)\sin(\phi), \\ z &= \cos(\theta). \end{aligned} \tag{5.1}$$

The Hilbert space of square-integrable functions on the sphere, $L^2(\mathbb{S}^2)$, is defined through the corresponding inner product. For two functions $f, g \in L^2(\mathbb{S}^2)$ we have

$$\langle f,\ g\rangle \stackrel{\text{def}}{=} \int_{\mathbb{S}^2} f(\xi)\overline{g(\xi)}\mathrm{d}\xi, \tag{5.2}$$

where $\mathrm{d}\xi = \sin(\theta)\,\mathrm{d}\theta\,\mathrm{d}\phi$ is the usual rotationally invariant measure on the sphere. With respect to this inner product, spherical harmonics form a natural orthonormal Fourier basis for $L^2(\mathbb{S}^2)$. They are defined as [65]

$$Y_\ell^m(\theta, \phi) = N_\ell^m P_\ell^{|m|}(\cos\theta)\mathrm{e}^{\mathrm{j}m\phi}, \tag{5.3}$$

where the normalization constant is

$$N_\ell^m = (-1)^{(m+|m|)/2}\sqrt{\frac{(2\ell+1)}{4\pi}\frac{(l-|m|)!}{(l+|m|)!}}, \tag{5.4}$$

and $P_\ell^m(x)$ is the associated Legendre polynomial of degree ℓ and order m. Note that different communities sometimes use different normalizations and sign conventions in the definitions of spherical harmonics and associated Legendre polynomials. As long as applied consistently, the choice of convention does not affect our results.[2]

In this paper, we adopt the following definition:

$$P_\ell^m(x) \overset{\text{def}}{=} (-1)^m(1-x^2)^{m/2}\frac{d^m}{dx^m}P_\ell(x), \text{ for } m \geqslant 0, \tag{5.5}$$

where $P_\ell(x)$ is the Legendre polynomial of degree ℓ [1].

Any square integrable function on the sphere, $f \in L^2(\mathbb{S}^2)$, can be expanded in the spherical harmonic basis,

$$f(\theta,\phi) = \sum_{\ell=0}^{\infty}\sum_{|m|\leqslant\ell}\widehat{f}_\ell^m Y_\ell^m(\theta,\phi), \tag{5.6}$$

where the Fourier coefficients are computed as

$$\widehat{f}_\ell^m = \langle f, Y_\ell^m\rangle = \int_{\mathbb{S}^2} f(\xi)\overline{Y_\ell^m(\xi)}\mathrm{d}\xi. \tag{5.7}$$

The coefficients $\left[\widehat{f}_\ell^m, (\ell,m)\in\mathcal{I}\right]$ form a countable set supported on an infinite triangle of indices,

$$\mathcal{I} = \left\{(\ell,m)\in\mathbb{Z}^2 \mid \ell \geqslant 0, |m| \leqslant \ell\right\}. \tag{5.8}$$

We say that f is *bandlimited* with bandwidth L if $\widehat{f}_\ell^m = 0$ for $\ell \geqslant L$. Often we think of L as the smallest integer such that this holds. For a bandlimited function, the triangle $\mathcal{I}$ is cut off at $\ell = L$. In what follows, we use

$$\mathcal{I}_L \overset{\text{def}}{=} \left\{(\ell,m)\in\mathbb{Z}^2 \mid 0 \leqslant \ell < L, |m| \leqslant \ell\right\} \tag{5.9}$$

to represent the spectral support of a bandlimited function with bandwidth L. The set $\mathcal{I}_L$ contains L^2 indices, so we can represent the spectrum as an L^2-dimensional column vector

$$\widehat{\boldsymbol{f}} \overset{\text{def}}{=} \left[\widehat{f}_0^0,\ \widehat{f}_1^{-1},\ \widehat{f}_1^0,\ \widehat{f}_1^1,\ \ldots,\ \widehat{f}_{L-1}^{-L+1},\ \ldots,\ \widehat{f}_{L-1}^{L-1}\right]^\top. \tag{5.10}$$

5.2.2 Rotations and Convolutions on the Sphere

Let $\mathbb{SO}_3$ denote the group of rotations in $\mathbb{R}^3$; any rotation $\varrho \in \mathbb{SO}_3$ is parameterized by three angles that specify rotations about three distinct axes. Thus we can write $\varrho = \varrho(\alpha,\beta,\gamma)$. The most common parameterization is called *Euler angles* [212].

---

[2] It is common to write the spherical harmonic order m in the superscript. We will keep this convention for the associated Legendre polynomials $P_\ell^{|m|}$, spherical harmonics Y_ℓ^m, normalization constants N_ℓ^m and the spherical Fourier coefficients $\widehat{f}_\ell^m$. It is not to be confused with integer powers such as x^ℓ.

A citeounter-clockwise rotation of a vector $\boldsymbol{x} \in \mathbb{R}^3$ about the z-axis is achieved by multiplying $\boldsymbol{x}$ by the corresponding rotation matrix,

$$\boldsymbol{R}_z(\alpha) = \begin{bmatrix} \cos\alpha & -\sin\alpha & 0 \\ \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

where α is the rotation angle. Rotation matrices around axes x and y can be defined analogously.

We use $\Lambda(\varrho)$ to represent the rotation operator corresponding to ϱ, that acts on spherical functions. Thus for f a function on the sphere, $\Lambda(\varrho)f$ represents the rotated function, defined as

$$[\Lambda(\varrho)f](\xi) \overset{\text{def}}{=} f(\varrho^{-1} \circ \xi), \tag{5.11}$$

where ρ^{-1} is the inverse rotation of ρ, and by $\varrho^{-1} \circ \xi$ we mean pre-multiplying by $\boldsymbol{R}(\varrho^{-1})$ the unit column vector corresponding to ξ, *cf.* (5.1). Compare this definition with the Euclidean case where shifting the argument to the left (subtracting a positive number) results in the shift of the function to the right.

There are various definitions of convolution on the sphere, all being non-commutative. One function, call it f, provides the weighting for the rotations of the other function h. A standard definition is then [65, 57]

$$\begin{aligned} [f * h](\xi) &\overset{\text{def}}{=} \left[\left(\frac{1}{2\pi} \int_{\mathbb{SO}_3} \mathrm{d}\varrho \cdot f(\varrho \circ \eta) \cdot \Lambda(\varrho) \right) h \right](\xi) \\ &= \frac{1}{2\pi} \int_{\mathbb{SO}_3} f(\varrho \circ \eta) h(\varrho^{-1} \circ \xi) \mathrm{d}\varrho, \end{aligned} \tag{5.12}$$

where $\eta \in \mathbb{S}^2$ is the north pole. It is easy to verify that this definition generalizes the standard convolution in Euclidean spaces [214], with the rotation operator ϱ playing the same role as translations do on the line. Because the spherical convolution is not commutative, it is important to fix the ordering of the arguments. In our case, the second argument—h in (5.12)—will always be the filter, *i.e.*, the observation kernel.

The familiar convolution–multiplication rule in standard Euclidean domains [214] holds for spherical convolutions too. It can be shown [65, Theorem 1] that for any two functions $f, h \in L^2(\mathbb{S}^2)$, the Fourier transform of their convolution is a pointwise product of the transforms, *i.e.*,

$$\widehat{(f * h)}_\ell^m = \sqrt{\frac{4\pi}{2\ell + 1}}\, \hat{f}_\ell^m \, \hat{h}_\ell^0. \tag{5.13}$$

We note that f can also be a generalized function (a distribution). In particular, we consider spherical Dirac delta functions, defined as [67]

$$\delta(\theta, \phi; \theta_0, \phi_0) = \frac{\delta(\theta - \theta_0)\delta(\phi - \phi_0)}{\sin(\theta)}, \tag{5.14}$$

and weighted sums of Dirac deltas. To lighten the notation, we often write $\delta(\xi; \xi_0)$. With the definition in (5.14), it is ensured that

$$\int_{\mathbb{S}^2} \delta(\xi; \xi_0) \mathrm{d}\xi = 1, \ \forall \xi_0 \in \mathbb{S}^2. \tag{5.15}$$

5.2.3 Problem Formulation

Consider a collection of K Diracs on the sphere,

$$f(\xi) = \sum_{k=1}^{K} \alpha_k \delta(\xi; \xi_k), \tag{5.16}$$

where the weights $\{\alpha_k \in \mathbb{C}\}_{k=1}^{K}$ and the locations of the Diracs $\{\xi_k = (\theta_k, \phi_k)\}_{k=1}^{K}$ are all unknown parameters. Let $y(\xi)$ be a filtered version of $f(\xi)$, *i.e.*,

$$y(\xi) = [f * h](\xi),$$

where the filter (or sampling kernel) $h(\xi)$ is a bandlimited function with bandwidth L. We further assume that the spherical Fourier transform of $h(\xi)$ is nonzero within its spectral support, *i.e.*, $\widehat{h}_\ell^m \neq 0$ for all $\ell < L$. Given spatial samples of $y(\xi)$, we would like to reconstruct $f(\xi)$, or equivalently, to recover the unknown parameters $\{(c_k, \xi_k)\}_{k=1}^{K}$.

Since the filtered signal $y(\xi)$ is bandlimited, we can use the sampling theorems for bandlimited functions on the sphere (*e.g.*, [65, 141]) or direct linear inversion (see Section 5.3.1) to recover its Fourier spectrum $\widehat{y}_\ell^m$ from its spatial samples of sufficient density. Using the convolution-multiplication identity in (5.13), we can then recover the lowpass subband of $f(\xi)$ as

$$\widehat{f}_\ell^m = [(2\ell+1)/(4\pi)]^{1/2} \cdot \left(\widehat{y}_\ell^m \,/\, \widehat{h}_\ell^0\right),$$

for $0 \leqslant \ell < L$ and $|m| \leqslant \ell$. Being a collection of Diracs, $f \notin L^2(\mathbb{S}^2)$, but its Fourier transform $\widehat{f}_\ell^m$ can still be computed via (5.7) in the sense of distributions as

$$\begin{aligned} \widehat{f}_\ell^m &= \sum_{k=1}^{K} \alpha_k \overline{Y_\ell^m(\theta_k, \phi_k)} \\ &= N_\ell^m \sum_{k=1}^{K} \alpha_k P_\ell^{|m|}(\cos\theta_k) \mathrm{e}^{-\mathrm{j} m \phi_k}. \end{aligned} \tag{5.17}$$

The problems we address in this chapter can now be stated as follows: Can we reconstruct a collection of K Diracs on the sphere from its Fourier coefficients $\widehat{f}_\ell^m$ in the lowpass subband $\mathcal{I}_L$ as defined in (5.9)? If so, then what is the minimum bandwidth L that allows us to do it? In practice, the sampling kernel is often given and not subject to our control. In this case, the previous question can be reformulated as determining the maximum number of spikes that we can reconstruct at a given bandwidth L.

5.3 Sampling Spherical FRI Signals

In this section we address the questions stated above. Our main result can be summarized in the following theorem:

Theorem 5.1

Let f be a collection of K Diracs on the sphere $\mathbb{S}^2$, with complex weights $\{\alpha_k\}_{k=1}^K$ at locations $\{\xi_k = (\theta_k, \phi_k)\}_{k=1}^K$, as in (5.16). Convolve f with a bandlimited sampling kernel h_L, where the bandwidth $L \geqslant K + \sqrt{K}$, and sample the resulting signal $[f * h_L](\xi)$ at L^2 points $\{\psi_n \in \mathbb{S}^2\}_{n=1}^{L^2}$ chosen uniformly at random on $\mathbb{S}^2$. Then almost surely the samples

$$f_n = [f * h_L](\psi_n), \quad n = 1, \ldots, L^2$$

are a sufficient characterization of f.

We provide a constructive proof of this theorem by presenting an algorithm that can efficiently recover K localized spikes from L^2 samples, where $L \geqslant K + \sqrt{K}$. Before presenting the algorithm and the proof, we first define some relevant notation and state two lemmas.

5.3.1 From Samples to the Fourier Transform

Our algorithms perform computation with spectral coefficients. In practice, we have access to spatial samples of the function, so we need a procedure to convert between the spatial and the Fourier representations. We first describe a method to compute the Fourier transform from samples taken at generically placed sampling points.

Let the function $f \in L^2(\mathbb{S}^2)$ have bandwidth L; then we can express it as

$$f(\theta, \phi) = \sum_{\ell=0}^{L-1} \sum_{m=-\ell}^{\ell} \widehat{f}_\ell^m Y_\ell^m(\theta, \phi). \tag{5.18}$$

Choose a set of sampling points $\{\psi_n \in \mathbb{S}^2\}_{n=1}^N$, and let $\boldsymbol{Y} = [y_{n,(\ell,m)}]$ where $y_{n,(\ell,m)} = Y_\ell^m(\psi_n)$. Furthermore, let $\boldsymbol{f} = [f(\psi_1), \ \ldots, \ f(\psi_n)]^\top$ be the vector of samples of f. We can then write

$$\boldsymbol{f} = \boldsymbol{Y}\widehat{\boldsymbol{f}}, \tag{5.19}$$

where $\widehat{\boldsymbol{f}}$ is the L^2-dimensional vector of spectral coefficients as defined in (5.10). The goal is to recover the spectral coefficients $\widehat{\boldsymbol{f}}$. We can recover $\widehat{\boldsymbol{f}}$ from $\boldsymbol{f}$ as soon as the matrix $\boldsymbol{Y}$ has full column rank. In that case, we compute

$$\widehat{\boldsymbol{f}} = \boldsymbol{Y}^\dagger \boldsymbol{f}, \tag{5.20}$$

where $\boldsymbol{Y}^\dagger$ denotes the Moore-Penrose pseudoinverse of the matrix $\boldsymbol{Y}$.

In particular, if we draw the samples uniformly at random on the sphere, we can show that $\boldsymbol{Y}$ is regular with probability one:

Proposition 5.1

Draw N sampling points from any absolutely continuous probability measure on the sphere (e.g. uniformly at random). Then $\boldsymbol{Y}$ has full column rank almost surely if $N \geqslant L^2$, that is, if it has at least as many rows as columns.

The proof of this proposition is identical to that of Theorem 3.2 in [11], and is thus omitted.

The above result indicates that we can recover the spectral coefficients $\widehat{f}_\ell^m$ in the lowpass region $\mathcal{I}_L$ from L^2 samples taken at generic points on the sphere. The reconstruction requires a matrix inversion as in (5.20).

Much faster reconstruction is possible when the function is sampled on certain regular grids. In that case, we can leverage the structure of $\boldsymbol{Y}$ to accelerate the matrix inversion. Such efficient schemes were proposed by Driscoll and Healy [65], requiring $4L^2$ samples; by McEwen and Wiaux [141], requiring $2L^2$ samples; and most recently, by Khalid, Kennedy and McEwen [107], requiring L^2 samples.

5.3.2 The Data Matrix

Using the definition of associated Legendre polynomials in (5.5), we rewrite the spherical harmonics (5.3) as

$$Y_\ell^m(\theta, \phi) = \widetilde{N}_\ell^m (\sin\theta)^{|m|} \left[\frac{d^{|m|}}{d(\cos\theta)^{|m|}} P_\ell(\cos\theta) \right] \mathrm{e}^{\mathrm{j}m\phi}, \tag{5.21}$$

where $\widetilde{N}_\ell^m = (-1)^m N_\ell^m$.

The essential observation is that the bracketed term in (5.21) is a polynomial in $x = \cos\theta$. At bandwidth L, the largest spherical harmonic degree is $L-1$, so the largest power of x in (5.21) is $L-1$ as well. It follows that we can rewrite the derivative term as a linear combination of powers of x, *i.e.*

$$\widetilde{N}_\ell^m \frac{d^{|m|}}{d(\cos\theta)^{|m|}} P_\ell(\cos\theta) = \boldsymbol{c}_{\ell m}^\top \boldsymbol{x}, \tag{5.22}$$

where $\boldsymbol{x} \overset{\text{def}}{=} [x^{L-1},\ x^{L-2},\ \cdots,\ x,\ 1]^\top$, $x = \cos\theta$ and $\boldsymbol{c}_{\ell m} \in \mathbb{R}^L$ contains the corresponding polynomial coefficients.

Using the dot-product formulation (5.22), the spectrum of f, as given by (5.17), can be expressed as

$$\widehat{f}_\ell^m = \boldsymbol{c}_{\ell m}^\top \sum_{k=1}^{K} \alpha_k \boldsymbol{x}_k (\sin\theta_k)^{|m|} \mathrm{e}^{-\mathrm{j}m\phi_k}, \tag{5.23}$$

where $\boldsymbol{x}_k \overset{\text{def}}{=} [x_k^{L-1},\ x_k^{L-2},\ \cdots,\ x_k,\ 1]^\top$ with $x_k = \cos\theta_k$, and we factored $\boldsymbol{c}_{\ell m}^\top$ out of the summation as it does not depend on k.

A key ingredient in our proposed algorithm is what we call the *data matrix* $\boldsymbol{\Delta}$, formed as a product of three matrices,

$$\boldsymbol{\Delta} \overset{\text{def}}{=} \boldsymbol{X}\boldsymbol{A}\boldsymbol{U}, \tag{5.24}$$

where

$$\boldsymbol{X} = [\boldsymbol{x}_1,\ \cdots,\ \boldsymbol{x}_K] \in R^{L\times K}, \tag{5.25}$$

is a Vandermonde matrix with roots $\cos\theta_k$, $\boldsymbol{A} = \operatorname{diag}(\alpha_1, \ldots, \alpha_K)$ is the diagonal matrix of Dirac magnitudes, and we define

$$\boldsymbol{U} = [u_{km}] \in \mathbb{R}^{K\times(2L-1)}, \tag{5.26}$$

with $u_{km} \overset{\text{def}}{=} (\sin\theta_k)^{|m|} e^{-\mathrm{j}m\phi_k}$.

It is convenient to keep a non-standard indexing scheme for the rows and columns of $\boldsymbol{\Delta}$, as illustrated in Figure 5.1B. Rows of $\boldsymbol{\Delta}$, indexed by p, correspond to decreasing powers of $\cos\theta_k$, from $p = L-1$ at the top, to $p = 0$ at the bottom; columns correspond to u_{km}, with m increasing from $-L+1$ on the left, to $L-1$ on the right. We see from (5.23) and (5.24) that computing any spectral coefficient $\widehat{f}_\ell^m$ amounts to applying a linear functional on $\boldsymbol{\Delta}$ as follows

$$\widehat{f}_\ell^m = \boldsymbol{c}_{\ell m}^\top \boldsymbol{\Delta} \boldsymbol{e}_m = \left\langle \boldsymbol{c}_{\ell m} \boldsymbol{e}_m^\top,\ \boldsymbol{\Delta} \right\rangle_F, \tag{5.27}$$

where $\boldsymbol{e}_m \in \mathbb{R}^{2L-1}$ is the vector with one in position m for $-L < m < L$, and zeros elsewhere, and $\langle \,\cdot\, , \,\cdot\, \rangle_F$ denotes the standard inner product between two matrices, defined as $\langle \boldsymbol{A},\ \boldsymbol{B} \rangle_F = \sum_{ij} \overline{a_{ij}} b_{ij} = \operatorname{trace}(\boldsymbol{A}^* \boldsymbol{B})$.

The last expresion in (5.27) implies that the spectral coefficient $\widehat{f}_\ell^m$ can be obtained as an inner product between the data matrix $\boldsymbol{\Delta}$ and a mask $\boldsymbol{c}_{\ell m} \boldsymbol{e}_m^\top$ that is overlaid over $\boldsymbol{\Delta}$. One can verify that the support of this mask for $\widehat{f}_\ell^m$ is on the column corresponding to m, and on the rows corresponding to $0 \leqslant p < L - |m|$. That means that certain parts of the data matrix are not involved in the creation of any spectral coefficient; consequently, they cannot be recovered from the spectrum. Nevertheless, we can recover a large part:

Lemma 5.1

There is a one-to-one linear mapping between the spherical harmonic coefficients in the lowpass subband, $\left[\widehat{f}_\ell^m, (\ell, m) \in \mathcal{I}_L\right]$, *and the triangular part of the data matrix* $\boldsymbol{\Delta}$ *indexed by* $\mathcal{J}_L = \{(p, m) \mid 0 \leqslant |m| \leqslant p < L\}$ *(with indexing as illustrated in Figure 5.1).*

Proof: It is straightforward to verify that all the masks $\boldsymbol{c}_{\ell m} \boldsymbol{e}_m^\top$ for $0 \leqslant |m| \leqslant \ell < L$ are supported on the triangular part of $\boldsymbol{\Delta}$, as indexed by $\mathcal{J}_L$. Because the number of such masks coincides with the number of entries in the triangular part, and no mask is identically zero, it only remains to show that the masks are linearly independent. For $m_1 \neq m_2$, this is true because their supports are disjoint ($\boldsymbol{e}_{m_1}^\top$ and $\boldsymbol{e}_{m_2}^\top$ activate different columns),

$$\operatorname{supp}(\boldsymbol{c}_{\ell_1 m_1} \boldsymbol{e}_{m_1}^\top) \cap \operatorname{supp}(\boldsymbol{c}_{\ell_2 m_2} \boldsymbol{e}_{m_2}^\top) = \varnothing \tag{5.28}$$

for any ℓ_1, ℓ_2. For $\ell_1 < \ell_2$ and $m_1 = m_2 = m$, $\boldsymbol{c}_{\ell_1 m}^\top \boldsymbol{x}$ and $\boldsymbol{c}_{\ell_2 m}^\top \boldsymbol{x}$ are polynomials of different degrees (*c.f.* (5.22)),

$$\deg(\boldsymbol{c}_{\ell_1 m}^\top \boldsymbol{x}) = (\ell_1 - |m|) < (\ell_2 - |m|) = \deg(\boldsymbol{c}_{\ell_2 m}^\top \boldsymbol{x}, \tag{5.29}$$

where $\deg(\,\cdot\,)$ denotes the degree of the polynomial in the argument. Therefore, $\operatorname{supp}(\boldsymbol{c}_{\ell_1 m} \boldsymbol{e}_m^\top) \neq \operatorname{supp}(\boldsymbol{c}_{\ell_2 m_2} \boldsymbol{e}_m^\top)$, and in particular $\boldsymbol{c}_{\ell_2 m}$ is linearly independent from all $\boldsymbol{c}_{\ell m}$ such that $\ell < \ell_2$. This implies that all masks are linearly independent. Thus the mapping

$$\begin{aligned} \boldsymbol{\Delta} \mapsto & \left[\left\langle \boldsymbol{c}_{\ell m} \boldsymbol{e}_m^\top,\ \boldsymbol{\Delta} \right\rangle_F,\ 0 \leqslant |m| \leqslant \ell < L \right] \\ & = \left[\widehat{f}_\ell^m, 0 \leqslant |m| \leqslant \ell < L \right] \end{aligned} \tag{5.30}$$

is one-to-one on $\mathcal{J}_L$. ■

5.3.3 Reconstruction by Generalized Annihilating Filtering

An element of the data matrix $\boldsymbol{\Delta}$ at the position (p, m) (with reference to Figure 5.1B) can be expanded as

$$d_{pm} = \sum_{k=1}^{K} \alpha_k x_k^p (\sin \theta_k)^{|m|} \mathrm{e}^{-\mathrm{j} m \phi_k}, \tag{5.31}$$

where p varies from 0 to $L-1$, and m from $-(L-1)$ to $(L-1)$. For either positive or negative m, the sum (5.31) is a sum of 2D exponentials. Lemma 5.1 implies that we can recover the shaded triangular part of the data matrix in Figure 5.1 from the spectrum. In what follows, we propose a new algorithm to recover the parameters of the Diracs from that triangular part.

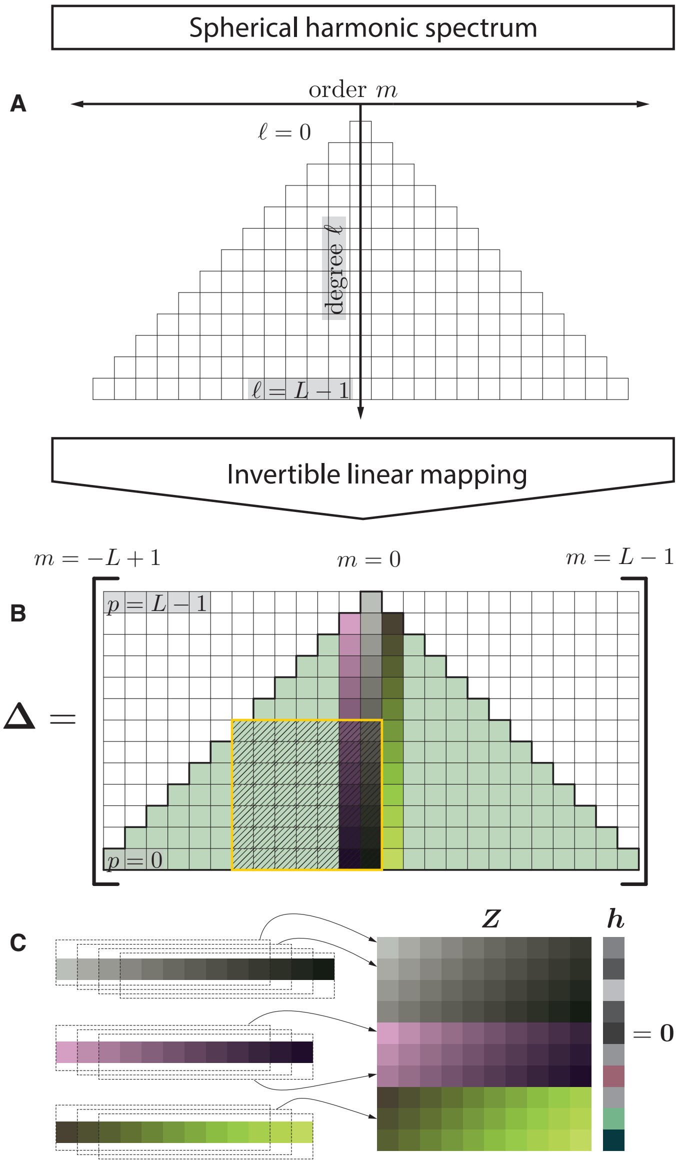


Figure 5.1: Illustration of Algorithm 5.1. Spherical harmonic spectrum (A) is linearly mapped onto the shaded triangular part of the data matrix $\boldsymbol{\Delta}$ (B). Columns of the data matrix are indexed from left to right by m, $-(L-1) \leqslant m \leqslant (L-1)$, corresponding to spherical harmonic order. Rows are indexed from bottom to top by p, $0 \leqslant p \leqslant (L-1)$ corresponding to powers of $\cos\theta$. Note that the triangular part of the data matrix *does not* coincide with the spherical harmonic spectrum, although there is a one-to-one linear mapping between the two (see Lemma 5.1). Existing results on 2D harmonic retrieval can exploit only a small part of the data matrix, for example the hatched square (see Section 5.3.4). Finally, sufficiently long columns of $\boldsymbol{\Delta}$ are rearranged in the block-Hankel-structured annihilation matrix $\boldsymbol{Z}$, whose nullspace contains exactly the sought annihilation filter, $\boldsymbol{h}$ (C).

The vector $\boldsymbol{d}_m \overset{\text{def}}{=} \boldsymbol{\Delta} \boldsymbol{e}_m$ is a linear combination of columns of $\boldsymbol{X}$, *i.e.*, it is a linear combination of K exponentials with bases x_k,

$$d_{pm} = \sum_{k=1}^{K} (\alpha_k u_{km}) x_k^p, \tag{5.32}$$

where $x_k = \cos(\theta_k)$. Similarly to the standard Euclidean FRI sampling [215], we can use the *annihilating filter* technique to estimate the roots $\{x_k = \cos\theta_k\}_{k=1}^K$ of these exponentials.

The annihilating filter is a finite impulse response (FIR) filter with zeros positioned so that it annihilates signals of the form (5.32). Consider an FIR filter $H(z)$ with the transfer function

$$H(z) \overset{\text{def}}{=} \prod_{k=1}^{K} (1 - x_k z^{-1}) \overset{\text{def}}{=} \sum_{n=0}^{K} h_n z^{-n}, \tag{5.33}$$

where $\boldsymbol{h} = [h_0,\ h_1,\ \ldots,\ h_K]^\top$ is the vector of filter coefficients. It holds that $\boldsymbol{h} * \boldsymbol{d}_m \equiv \boldsymbol{0}$ (see Appendix 5.A) for any m, provided that $\boldsymbol{d}_m$ is of length at least $K + 1$. Equivalently,

$$[d_{n,m},\ d_{n-1,m},\ \ldots,\ d_{n-K,m}] \begin{bmatrix} h_0 \\ h_1 \\ \vdots \\ h_K \end{bmatrix} = 0, \tag{5.34}$$

for $n \geqslant K$. In our scenario, we do not know the bases of the exponentials $\{x_k\}_{k=1}^K$—they are exactly the parameters we aim to estimate. Thus we do not know the filter $H(z)$ either.

Up to a scaling factor, there is a unique $(K + 1)$-tap filter $H(z)$ with the sought property. The orthogonality relation (5.34) says that $\boldsymbol{h}$ lives in the nullspace of $[d_{n,m}\ d_{n-1,m}\ \cdots\ d_{n-K,m}]$; we need at least K such vectors to make their joint nullspace one-dimensional, thus to pinpoint $\boldsymbol{h}$. Once the filter coefficients are found, we can obtain the unknown parameters $\{x_k\}$ by root finding and using the factorization in (5.33).

For the annihilating filter technique to be applicable, we need to ensure that all the colatitude angles θ_k are distinct. Furthermore, the form of our equations reveals that for $\theta_k \in \{0, \pi\}$, $u_{km} = 0$ for all m. In the parameterization (5.24), this is equivalent to setting $\alpha_k = 0$, and it prevents us from recovering the corresponding Dirac. This behavior is undesirable, but we can guarantee that no Dirac sits on a pole by first applying a random rotation. This fact is formalized in the following lemma, which follows immediately from the absolute continuity of the Haar measure.

Lemma 5.2

Consider a collection of Dirac delta functions on the sphere, $f(\xi) = \sum_{k=1}^K \alpha_k \delta(\xi; \xi_k)$, and a random rotation ϱ drawn from the Haar measure on $\mathbb{SO}_3$ (i.e. uniformly over the elements of the group). Then with probability 1, $\Lambda(\varrho) f$ contains Diracs with distinct colatitude angles, $\theta_i \neq \theta_j$ for $i \neq j$, and no Dirac is on the pole, $\theta_k \notin \{0, \pi\}$ for all k.

We are now well-equipped to prove the main result.

Proof of Theorem 5.1 : We provide a constructive proof, summarized in Algorithm 5.1. First observe that L^2 random samples almost surely suffice to compute the spectral coefficients $\widehat{f}_\ell^m$ in the lowpass subband $\mathcal{I}_L$ with bandwidth L, as detailed in Section 5.3.1 (see Proposition 5.1). By Lemma 5.1, we can then compute the shaded part of $\boldsymbol{\Delta}$ given the spectrum $\widehat{\boldsymbol{f}}$.

Our aim is to construct the annihilating matrix $\boldsymbol{Z}$, structured as follows

$$\boldsymbol{Z} = \begin{bmatrix} d_{L-1,0} & d_{L-2,0} & \cdots & d_{L-K-1,0} \\ d_{L-2,0} & d_{L-3,0} & \cdots & d_{L-K-2,0} \\ \vdots & \vdots & & \vdots \\ d_{K,0} & d_{K-1,0} & \cdots & d_{0,0} \\ d_{L-2,1} & d_{L-3,1} & \cdots & d_{L-K-2,1} \\ d_{L-3,1} & d_{L-4,1} & \cdots & d_{L-K-3,1} \\ \vdots & \vdots & & \vdots \end{bmatrix}. \tag{5.35}$$

$\boldsymbol{Z}$ is constructed by stacking segments of length $(K+1)$ extracted from the columns of $\boldsymbol{\Delta}$. From the annihilation property (5.34), it follows that the nullspace of $\boldsymbol{Z}$ contains the sought annihilating filter.

The trick now is to count how many such segments we can get from the shaded part of $\boldsymbol{\Delta}$. For $m = 0$, p varies from 0 to $L-1$. Therefore, we can construct $L-K$ rows of the matrix $\boldsymbol{Z}$. For $m = 1$, p varies from 0 to $L-2$, so we can construct $L-K-1$ rows of $\boldsymbol{Z}$, and the same goes for $m = -1$. This process is illustrated in Figures 5.1B and 5.1C. Summing up, we get the total number of rows of $\boldsymbol{Z}$ that we can construct from the available spectrum,

$$\begin{aligned} \# &= (L-K) + 2 \times (L-K-1) + \cdots + 2 \times 1 \\ &= (L-K)^2. \end{aligned} \tag{5.36}$$

$\boldsymbol{Z}$ needs at least K rows, as we need a 1D nullspace. Thus

$$\begin{aligned} (L-K)^2 &\geqslant K \\ \Rightarrow L &\geqslant K + \sqrt{K}. \end{aligned} \tag{5.37}$$

In Appendix 5.C we show that $\boldsymbol{Z}$ has rank K as soon as it has K or more rows. In other words, it has a one-dimensional nullspace, and thus the annihilating filter coefficients are uniquely determined, up to a scaling factor.

We find the parameters $\{\theta_k\}_{k=1}^K$ by taking the arc cosine of the roots of $H(z)$. This procedure is well-posed because arc cosine is one-to-one on $[0, \pi]$. To ensure that the roots are distinct, we apply a random rotation before the estimation, and the inverse of this random rotation after recovering all the parameters of the Diracs (invoking Lemma 5.2).

In order to recover the azimuths $\{\phi_k\}_{k=1}^K$, note that after recovering the colatitudes, we can construct the matrix $\boldsymbol{X}$, and compute $\boldsymbol{A}\boldsymbol{U}\boldsymbol{e}_m$ for $|m| \leqslant L-K$. The azimuths are then given as the phase difference between $\boldsymbol{A}\boldsymbol{U}\boldsymbol{e}_0$ and $\boldsymbol{A}\boldsymbol{U}\boldsymbol{e}_1$. The magnitudes α_k are obtained simply as $\boldsymbol{A}\boldsymbol{U}\boldsymbol{e}_0$. ■

5.3.4 Sampling Efficiency and Relation to Prior Work

Our proposed sampling scheme and the spherical FRI sampling theorem by Deslauriers-Gauthier and Marziliano [47] are both naturally expressed in terms of the bandwidth L of the sampling kernel required to recover K Diracs. In our case, the bandwidth requirement is that it be at least $K + \sqrt{K}$. This implies that we need at least $(K + \sqrt{K})^2$ spatial samples in order to recover the K Diracs. For comparison, the FRI sampling theorem of Deslauriers-Gauthier and Marziliano [47] requires $L \geqslant 2K$, and thus their algorithm recovers K Diracs given $4K^2$ samples. This is asymptotically four times the number of samples required by Algorithm 5.1.

Algorithm 5.1 Spherical Sparse Sampling

```
Input: Spatial samples of $f \in L^2(\mathbb{S}^2)$ with bandwidth $L$, number of Diracs $K$
Output: Colatitudes, azimuths and magnitudes $\{(\alpha_k, \theta_k, \phi_k)\}_{k=1}^{K}$ of the $K$ Diracs
Sample a random rotation $\varrho \sim \mathrm{Haar}(\mathbb{SO}_3)$
Apply $\varrho$ to $f$, $f \leftarrow \Lambda(\varrho)f$ (relabel sampling points)
Compute the spectrum $\widehat{f}$ from the rotated samples of $f$
Form $\boldsymbol{\Delta}$ from $\widehat{f}$ using the inverse mapping of (5.30)
Form $\boldsymbol{Z}$ from $\boldsymbol{\Delta}$ according to (5.35)
$\boldsymbol{h} \leftarrow$ Right singular vector of $\boldsymbol{Z}$ for smallest sing. val.
Compute the colatitudes, $(\theta_k)_{k=1}^{K} \leftarrow \arccos[\mathrm{Roots}(\boldsymbol{h})]$
Construct $\boldsymbol{X}$ from $x_k = \cos\theta_k$ according to (5.25)
Using $\boldsymbol{X}$ in (5.24), compute $\boldsymbol{A}\boldsymbol{U}\boldsymbol{e}_0$ and $\boldsymbol{A}\boldsymbol{U}\boldsymbol{e}_1$
$(\phi_k)_{k=1}^{K} \leftarrow \mathrm{Angle}\big[(\boldsymbol{A}\boldsymbol{U}\boldsymbol{e}_0) \oslash (\boldsymbol{A}\boldsymbol{U}\boldsymbol{e}_1)\big]$        ▷ See the note
$(\alpha_k)_{k=1}^{K} \leftarrow \boldsymbol{A}\boldsymbol{U}\boldsymbol{e}_0$
Apply the inverse of $\varrho$, $\xi_k = (\theta_k, \phi_k) \leftarrow \varrho^{-1} \circ \xi_k$, $\forall k$
```

▷ Note: we use the symbol $\oslash$ to denote the entrywise division of vectors.

The difference in sampling efficiency can be explained by the spectrum usage. Figure 5.2 illustrates the portion of the spectrum used by the two algorithms. We can see that the proposed algorithm is more efficient in that it uses a larger portion of the available spectrum to reconstruct the Diracs.

Similar problems have been considered in the literature on 2D harmonic retrieval [211]. However, these earlier works assume that the entire data matrix is known. In our case, $\boldsymbol{\Delta}$ is known only partially, as illustrated in Figure 5.1B. To apply the existing results on 2D harmonic retrieval, we could use a square portion that falls strictly inside a half of the triangle, either for $m \geqslant 0$ or for $m \leqslant 0$. However, we can see in Figure 5.1B that this is an inefficient use of available spectrum, and it requires an unnecessarily high sampling density.

As mentioned earlier, in most situations we do not get to choose L as it is fixed by the underlying physical process. Then the question is how many Diracs we can reconstruct given a kernel with a fixed bandwidth L. By solving $L \geqslant K + \sqrt{K}$ for K, we get that

$$K \leqslant L - (L + \tfrac{1}{4})^{1/2} + \tfrac{1}{2}. \tag{5.38}$$

In contrast, the algorithm in [47] can reconstruct up to $K = L/2$ Diracs.

5.3.5 Denoising Strategies

Theorem 5.1 and Algorithm 5.1 provide a tool to recover sparse signals on the sphere in the noiseless case. We may apply several procedures to improve the robustness of the algorithm in the presence of noise.

In general, if the samples are noisy then the annihilating matrix $\boldsymbol{Z}$ in (5.35) will not have a nontrivial nullspace. A simple and robust approach is to use the right singular vector corresponding to the smallest singular value of $\boldsymbol{Z}$ as the annihilation filter. Let $\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^H$ be the SVD of $\boldsymbol{Z}$; then we set $\boldsymbol{h} = \boldsymbol{v}_{K+1}$.

To further improve the algorithm performance, we can use the output of Algorithm 5.1 to initialize a local search for the minimizer of the ℓ^2 error between the spectrum generated by the

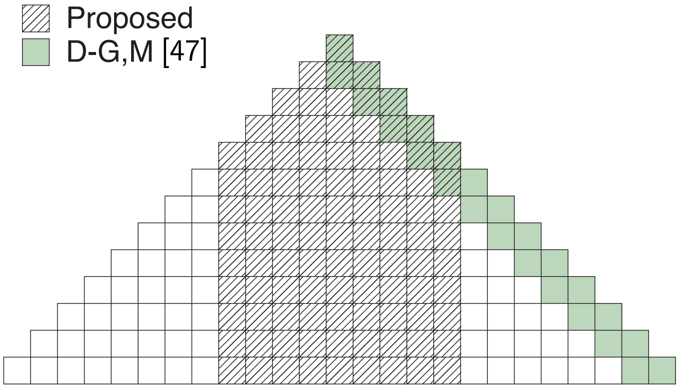


Figure 5.2: Spectrum usage for different algorithms. Spectral coefficients used by our algorithms are shown hatched. Spectrum used by the algorithm of Deslauriers-Gauthier and Marziliano [47] is shaded green. In the example, the bandwidth is set to $L = 12$, so the maximum number of Diracs that can be recovered by Algorithm 5.1 is $K = 9$. The algorithm in [47] recovers $K = 6$ Diracs.

estimated Diracs, and the measured spectrum,

$$\underset{(\widetilde{\alpha}_k, \widetilde{\theta}_k, \widetilde{\phi}_k)_{k=1}^K}{\text{minimize}} \sum_{\ell=0}^{L-1} \sum_{m=-\ell}^{\ell} \left| \widehat{f}_\ell^m - \sum_{k=1}^K \widetilde{\alpha}_k Y_\ell^m(\widetilde{\theta}_k, \widetilde{\phi}_k) \right|^2. \tag{5.39}$$

We note that directly solving (5.39) with a random starting point is hopeless due to a multitude of local minima.

5.3.6 Cramér-Rao Lower Bound

We evaluate the Cramér-Rao lower bound (CRLB) for the estimation problem. For simplicity we treat the $K = 1$ case, so that the minimal bandwidth is $L = 2$, and $\ell \in \{0, 1\}$. We assume that the spatial samples are taken on the sampling grid defined by McEwen–Wiaux [141], given at this bandwidth as

$$\begin{bmatrix} \boldsymbol{\theta} \\ \boldsymbol{\phi} \end{bmatrix} = \begin{bmatrix} \pi/3 & \pi/3 & \pi/3 & \pi & \pi & \pi \\ 0 & 0 & 2\pi/3 & 2\pi/3 & 4\pi/3 & 4\pi/3 \end{bmatrix}. \tag{5.40}$$

The resulting expressions for the elements of the Fisher information matrix are complicated, and there is no need to exhibit them explicitly. We give the details of the computation in Appendix 5.B, and we compute the CRLB numerically. The resulting bound, together with the MSE achieved by Algorithm 5.1 followed by the descent (5.39), is plotted in Figure 5.3 for two different spike colatitudes. As pointed out before, because our scheme is coordinate-system-dependent, the bound depends on the colatitude of the Dirac.

5.4 Applications

To showcase the versatility of the proposed algorithm, we present three stylized applications: 1) sampling diffusion processes on the sphere, 2) shot noise removal, and 3) sound source localization with spherical microphone arrays.

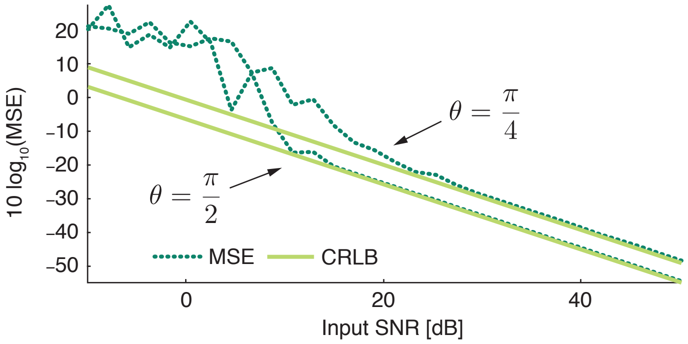


Figure 5.3: Comparison between the mean squared error (MSE) of the proposed algorithm in estimating the spherical location (θ, ϕ), with $K = 1$, and the Cramér-Rao lower bound (CRLB), at two different colatitudes. Note that the bound is different for different colatitudes of the spike, due to parameterization dependence. MSE is shown for the output of Algorithm 5.1 followed by the minimization of (5.39) using Matlab's `fminsearch` function.

5.4.1 Sampling Diffusion Processes on the Sphere

The diffusion process models many natural phenomena. Examples include heat diffusion and plume spreading from a smokestack. Often, the source of the diffusion process is localized in space, and instantaneous in time. Sampling such processes in Euclidean domains has been well studied [58, 131, 130].

Diffusion processes on the sphere are governed by the equation [28]

$$k\Delta v(\xi, t) = \frac{\partial}{\partial t} v(\xi, t), \tag{5.41}$$

where Δ is the Laplace-Beltrami operator on $\mathbb{S}^2$, and k is the diffusion constant. In the spherical harmonic domain, this becomes

$$-k\ell(\ell+1)\widehat{v}_\ell^m(t) = \frac{\partial}{\partial t}\widehat{v}_\ell^m(t), \tag{5.42}$$

giving the solution

$$\widehat{v}_\ell^m(t) = \mathrm{e}^{-\ell(\ell+1)kt}\widehat{v}_\ell^m(0), \tag{5.43}$$

where $\widehat{v}_\ell^m(0)$ is the spectrum of the initial distribution. Therefore, we interpret the term $\mathrm{e}^{-\ell(\ell+1)kt}\delta_{m,0}$ as the spectrum of the Green's function of the spherical diffusion equation. In other words, it is the spectrum of the diffusion kernel on the sphere. Then (5.43) should be interpreted as the convolution between the kernel and the initial distribution.

We consider the case when the diffusion process is initiated by K sources localized in space and time, *i.e.*, the initial distribution in (5.43) is

$$v(\xi; 0) = \sum_{k=1}^{K} \alpha_k \delta(\xi; \xi_k). \tag{5.44}$$

We show how to use the proposed sampling algorithm to estimate the locations and the strengths of the sources from spatial samples of the diffusion field taken at a later time t_0. Recovering all

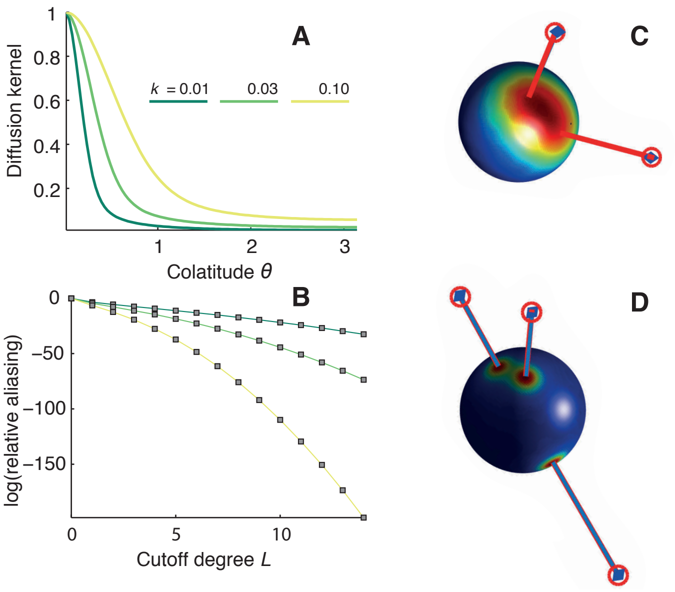


Figure 5.4: Estimating the release locations and magnitudes of diffusive sources on the sphere. We assume that the diffusive sources appear at time $t = 0$ s, and that we sample the field at time $t = 1$ s. Shape of the diffusion kernel as a function of θ is shown in subfigure A for three different values of the coefficient k (in units of inverse time). The logarithm of the aliasing error (5.47) is plotted as a function of the cutoff degree L in subfigure B. Subfigures C and D show a typical reconstruction result for $k = 0.1$ (2 sources) and $k = 0.01$ (3 sources). Magnitudes of the sources are represented by the distance of the corresponding symbols from the sphere's center. Blue diamonds represent true source locations and magnitudes, while red circles represent estimated source locations and magnitudes. The sphere color corresponds to the value of the function induced on the sphere by the sources (red is large, blue is small). Signal-to-noise ratio in both C and D was set to 30 dB. We used the approximate bandwidth of $L = 7$, so that the number of samples taken in either case was 49.

parameters (locations, amplitudes and release times) of multiple diffusion sources is a challenging task [58]. To focus on the proposed sampling result, we make the simplifying assumption that the K sources are released simultaneously, and at a known time ($t = 0$). In principle, the more challenging case of unknown and different release times can be handled by adapting the techniques derived in [131, 130], but these generalizations are out of the scope of this work.

In the spatial domain, the diffusion kernel at time t_0 after the release is given as

$$h_{\mathrm{dif}}(\xi; t_0) = \sum_{\ell=0}^{\infty} \mathrm{e}^{-\ell(\ell+1)kt_0} Y_\ell^0(\xi). \tag{5.45}$$

Combining (5.45) with (5.43) and the spherical convolution-multiplication rule (5.13), we get

$$\begin{aligned} v(\xi; t_0) &= v(\xi; 0) * h_{\mathrm{dif}}(\xi; t_0) \\ &= \sum_{k=1}^{K} c_k [\Lambda(\varrho_k) h_{\mathrm{dif}}(\,\cdot\,; t_0)](\xi). \end{aligned} \tag{5.46}$$

This signal is a sum of rotations of a known template. The diffusion kernel in (5.45) is not exactly bandlimited, but it is approximately so. We can therefore apply the spherical FRI theory and Algorithm 5.1 to recover the locations and the magnitudes of the diffusive sources.

Figure 5.4A shows the shape of the symmetric diffusion kernel as a function of the colatitude θ. The high degree of smoothness is reflected in an approximately bandlimited spectrum. This is demonstrated in Figure 5.4B, where we see that the aliasing energy due to spectral truncation, defined as

$$\varepsilon(L) = \frac{1}{\|v\|_2^2} \sum_{\ell=L}^{\infty} \frac{|\widehat{v}_\ell^m|^2}{2\ell+1}, \tag{5.47}$$

rapidly becomes negligible as we increase the cutoff bandwidth L. Figures 5.4C and 5.4D demonstrate accurate reconstruction of the localized diffusion sources at two different values of the diffusion coefficient (the detailed parameters of the numerical experiment are given in the figure caption).

5.4.2 Shot Noise Removal

Suppose that we sample a bandlimited function on the sphere, but a small number of samples are corrupted—they contain *shot noise*—due to sensor malfunction. Moreover, the identities of the malfunctioning sensors are not known *a priori*. Can we detect and correct these anomalous measurements? We show that our sampling results can be applied to solve this problem, provided that the number of erroneous sensors is not too large and that the original sampling grid is *oversampling* the bandlimited function. A similar idea was used in [139] to remove shot noise in the 1D Euclidean case.

For this application we assume that the samples are taken on a uniform grid on the sphere, $\{(\theta_p, \phi_q) \mid p, q \in \mathbb{Z}, 0 \leqslant p < 2L', 0 \leqslant q < 2L'\}$, defined by

$$\theta_p = \frac{p\pi}{2L'} \quad , \quad \phi_q = \frac{q\pi}{L'}. \tag{5.48}$$

Imagine now that we sample f on this sampling grid. Some samples are corrupted, so we measure $g(\theta_p, \phi_q) = f(\theta_p, \phi_q) + s_{pq}$, where

$$s_{pq} = \begin{cases} \text{nonzero} & (p,q) \in \mathcal{S} \\ \text{zero} & \text{otherwise,} \end{cases} \tag{5.49}$$

and $\mathcal{S}$ holds the indices of the corrupted samples. We will leverage an elegant quadrature rule by Driscoll and Healy [65]:

Theorem 5.2 (*[65, Theorem 3]*)

Let f be a bandlimited function on $\mathbb{S}^2$ such that $\widehat{f}_\ell^m = 0$ for $\ell \geqslant L'$. Then for $(\ell, m) \in \mathcal{I}_{L'}$ we have

$$\widehat{f}_\ell^m = \sum_{p=0}^{2L'-1} \sum_{q=0}^{2L'-1} a_p^{(L')} f(\theta_p, \phi_k) \overline{Y_\ell^m(\theta_p, \phi_q)}, \tag{5.50}$$

where the weights $a_p^{(L')}$ are defined in [65].

In other words, the Fourier coefficients $\widehat{f}_\ell^m$ can be expressed as a dot-product between the weighted sample values and the basis functions evaluated at the sampling points. In analogy with the Euclidean case, we now observe that the lowpass portion of the spectrum of f coincides with the lowpass portion of the spectrum of the generalized function obtained by placing weighted Diracs at grid points. Let f be bandlimited so that $\widehat{f}_\ell^m = 0$ for $\ell \geqslant L$. Let further $L < L'$; that is, the grid (5.48) oversamples f. Then the spectral coefficients can be expressed as the following inner product,

$$\widehat{f}_\ell^m = \left\langle \sum_{p,q=0}^{2L'-1} a_p^{(L')} f(\theta_p, \phi_q) \delta_{\theta_j,\phi_q},\ Y_\ell^m \right\rangle, \tag{5.51}$$

for $\ell < L$, $|m| \leqslant \ell$.

This is the key insight. Notice that the lowpass portion of the spectrum of g (for $\ell < L'$) can be written as

$$\widehat{g}_\ell^m = \widehat{f}_\ell^m + \left\langle \sum_{(p,q)\in\mathcal{S}} a_p^{(L')} s_{pq} \delta_{\theta_p,\phi_q}, Y_\ell^m \right\rangle. \tag{5.52}$$

But $\widehat{f}_\ell^m = 0$ for $\ell \geqslant L$, so the portion of the spectrum for $L \leqslant \ell < L'$ contains only the influence of the corrupted samples,

$$\widehat{g}_\ell^m = \left\langle \sum_{(p,q)\in\mathcal{S}} a_p^{(L')} s_{pq} \delta_{\theta_p,\phi_q}, Y_\ell^m \right\rangle, \quad L \leqslant \ell < L'. \tag{5.53}$$

Consequently, we can use this part of the spectrum to learn which samples are corrupted, and by how much. This is the subject of the following proposition.

Proposition 5.2

Let f be a signal on the sphere of bandwidth L. Then we can perfectly reconstruct f from corrupted samples taken on the grid (5.48), *as long as the number of corruptions K satisfies*

$$K \leqslant L' - L - \sqrt{L' - L + 1} + 1. \tag{5.54}$$

Proof: As discussed in Section 5.3, we can use any $m =$ const. line in the spectrum to get the rows of the annihilation matrix. However, we first need to compute the corresponding columns of the data matrix. From Figure 5.5, we see that the middle columns cannot be used for shot noise removal: we seek columns influenced only by corruptions. But the middle columns of the data matrix are obtained from the middle spectral columns (for $m < L$), so they are influenced both by the desired signal and the corruptions. This means that we can only use spectral bins for $L \leqslant m < L'$, as illustrated in Figure 5.5. For $m = L$ and $m = -L$, the number of segments of length $K+1$ that we can get is $L' - L - K$. For $m = L+1$ and $m = -(L+1)$ it is $L' - L - K - 1$, and so on. Summing up we have that the total number of consecutive segments of length $K+1$ we can use is

$$\begin{aligned} \# &= 2(L' - L - K) + 2(L' - L - K - 1) + \cdots + 2 \cdot 1 \\ &= (L' - L - K)(L' - L - K + 1). \end{aligned}$$

We need this number to be at least K, because we need K rows in the annihilation matrix. We thus obtain the claim of the proposition by solving the inequality $\# \geqslant K$. ■

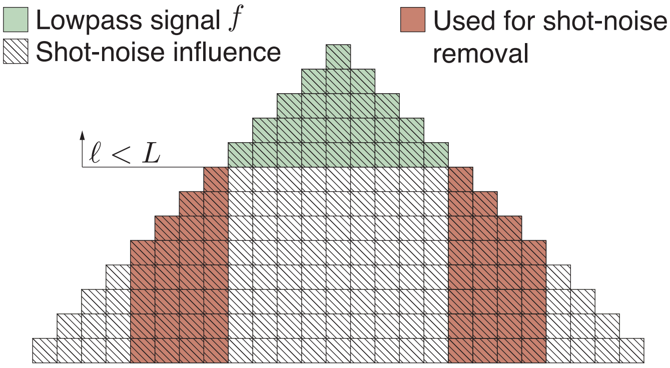


Figure 5.5: Spectrum structure in shot noise removal. Green-shaded bins get contribution from the desired signal f with bandwidth L; hatched bins are influenced by the shot noise; red-shaded columns are (i) long enough to annihilate shot noise and (ii) recoverable from the corrupted spectrum.

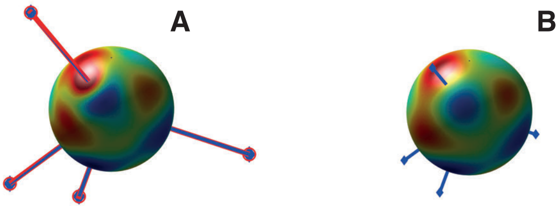


Figure 5.6: Shot noise removal via spherical FRI, for $L = 6$, $L' = 12$ and $K = 4$ malfunctioning sensors. Corrupted signal is shown in subfigure A, together with the true corruption values (blue diamonds) and the estimated corruptions (red circles); same signal with the shot noise removed is shown in B, with the correct sample values at the corrupted locations denoted by blue diamonds.

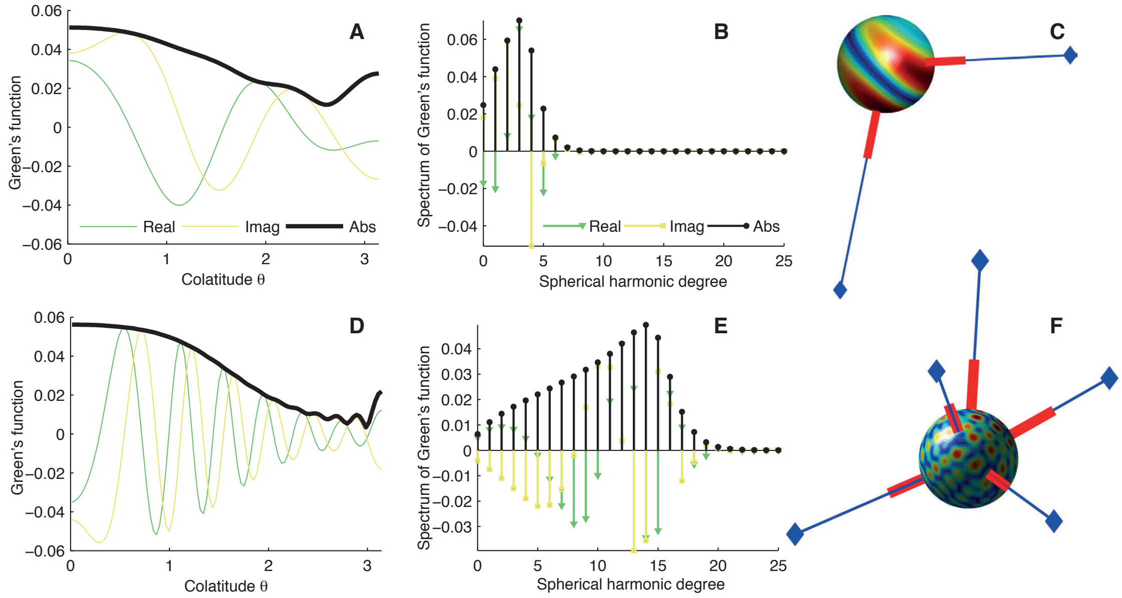


Figure 5.7: Multiple DOA estimation by a spherical microphone array. First row of subfigures corresponds to $f_1 = 1000$ Hz, and second row to $f_2 = 4000$ Hz. Sphere has a radius $r = 0.2$ m, and the source is located at $[0,\ 0,\ 3]^\top$ m. The real and imaginary parts, and the absolute value of the Green's function are shown in subfigures A and D. Real part, imaginary part and absolute value of the spectrum are shown in subfigures B and E. Subfigures C and F show the simulation results for $K = 2$ and $K = 5$, and random source placement. Blue diamonds represent the source locations, and thick red lines show the estimated directions. Size of the sphere is exaggerated for the purpose of illustration. The sphere color corresponds to the absolute value of the function induced on the sphere by the sources (microphones measure samples of this function). The bandwidth was set to $L = 12$ at 1000 Hz and to $L = 30$ at 4000 Hz.

After detecting the corrupted readings, we can use the estimated corruption values to estimate the function. Another option is to simply ignore them altogether, as we have more samples than the minimum number thanks to oversampling. A shot noise removal experiment is illustrated in Figure 5.6.

5.4.3 Sound Source Localization

Spherical microphone arrays measure a time-varying spherical signal. If the signal is induced by a collection of point sources, we can use the proposed spherical FRI sampling scheme to estimate the directions-of-arrival (DOAs) of the sources. For simplicity, we consider the narrowband case, *i.e.*, the sources emit a single sinusoid.

How does this example fit into our sparse sampling framework? In spherical microphone arrays, the microphones are distributed on the surface of a sphere, either open or rigid [98]. Therefore, the microphone signals represent samples of a time-varying function on $\mathbb{S}^2$. If a sound source emits a sinusoid, every microphone measures the amplitude and the phase of that sinusoid shaped by the characteristics of the propagating medium and of the spherical casing.

Equivalently, for every microphone we get a complex number.

Suppose that a source of unit intensity is located at $\boldsymbol{s}$, and that the microphones are mounted on a rigid sphere of radius r centered at the origin. The response measured by the microphone at $\boldsymbol{r}$, such that $\|\boldsymbol{r}\| = r$, is given by the corresponding Green's function. For a wavenumber $\kappa = 2\pi\nu/c$, where ν is the frequency and c is the speed of sound, the Green's function is [98]

$$g(\boldsymbol{r}|\boldsymbol{s},\kappa) = \frac{\mathrm{j}k}{4\pi}\sum_{\ell=0}^{\infty} b_\ell(\kappa r) h_\ell^{(1)}(\kappa s)(2\ell+1)P_\ell(\cos\alpha_{\boldsymbol{rs}}), \tag{5.55}$$

where $h_\ell^{(1)}$ is the spherical Hankel function of the first kind and of order ℓ, P_ℓ is the ℓth Legendre polynomial, and $\cos\alpha_{\boldsymbol{rs}} = \frac{1}{rs}\langle \boldsymbol{r}, \boldsymbol{s}\rangle$. Mode strength $b_\ell(kr)$ is defined as

$$b_\ell(\kappa r) \overset{\text{def}}{=} j_\ell(\kappa r) - \frac{j_\ell'(\kappa r)}{h_\ell^{(1)\prime}(\kappa r)} h_\ell^{(1)}(\kappa r), \tag{5.56}$$

where j_ℓ is the spherical Bessel function[3] of order ℓ, and prime $(\,\cdot\,)'$ denotes the derivative with respect to the argument.

The Green's function g should be seen as a filter that describes how the point source's influence spreads over the sphere. It is shown for two different frequencies in Figures 5.7A and 5.7D, while the corresponding spectra are given in Figures 5.7B and 5.7E. We see that the absolute pressure on the sphere has a similar shape for both frequencies, but the real and imaginary parts vary faster at higher frequencies, implying higher bandwidth. In both cases we observe that the Green's function is approximately bandlimited.[4]

Assume now that there are K sound sources at locations $\{\boldsymbol{s}_k\}_{k=1}^K$, with complex intensities $\{\alpha_k\}_{k=1}^K$. The resulting measurement by a microphone at point $\boldsymbol{r}$ is

$$f(\boldsymbol{r}) = \sum_{k=1}^{K} \alpha_i g(\boldsymbol{r}|\boldsymbol{s}_k,\kappa). \tag{5.57}$$

If all the source locations $\boldsymbol{s}_k$ are at the same distance from the sphere, then the Green's function (5.55) depends only on the angle between $\boldsymbol{r}$ and $\boldsymbol{s}$. For some fixed source distance d, we can define $h_{\mathrm{SSL}}(\xi) \overset{\text{def}}{=} g(\boldsymbol{x}_\xi r|\boldsymbol{x}_\eta d,\kappa)$, where $\boldsymbol{x}_\xi$ denotes the unit vector corresponding to ξ, $\boldsymbol{x}_\eta$ the unit vector corresponding to the north pole η, and the subscript SSL stands for *sound source localization.* Then (5.57) corresponds to a weighted sum of K rotations of a known template function h_{SSL},

$$f(\xi) = \sum_{k=1}^{K} \alpha_k h_{\mathrm{SSL}}(\varrho_k^{-1}\circ\xi). \tag{5.58}$$

As it is unrealistic to assume that the sources are all at the same distance, we hope that the shape of $g(\boldsymbol{r}|\boldsymbol{s},\kappa)$ does not (strongly) depend on $\|\boldsymbol{s}\|$. Indeed, it turns out that the shape is approximately preserved within a certain range, as illustrated in Figure 5.8. We therefore suppress the dependency of g on $\|\boldsymbol{s}\|$ and approximate (5.57) as follows,

---

[3] We use the standard symbol j_ℓ for the spherical Bessel function. Note the subtle difference from the imaginary unit j.

[4] This can be related to the spectral support of the plenacoustic function [3].

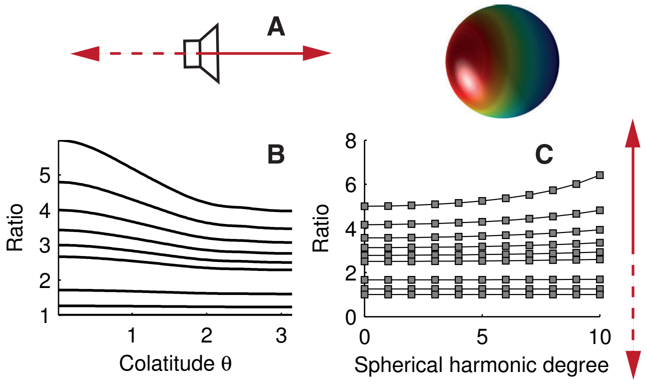


Figure 5.8: Ratios of Green's functions. We computed the Green's function for nine different source distances (1.0 m, 1.2 m, 1.4 m, 1.6 m, 1.8 m, 2.0 m, 3.0 m, 4.0 m, 5.0 m). Then we plotted the magnitude of the ratio of the Green's function at each distance and the Green's function at the largest distance (5 m), both in space (B) and in the spectrum (C). The more parallel the ratio curve is with the abscissa axis, the more similar the Green's function at that distance is to the Green's function at 5 m. Curves are plotted in the order of increasing distance in the direction of the dashed arrow (up to down), as indicated in (A).

$$\begin{aligned} f(\xi) &= \sum_{k=1}^{K} \alpha_k g(\xi | \boldsymbol{s}_k, \kappa) \\ &\approx \sum_{k=1}^{K} \widetilde{\alpha}_k h_{\mathrm{SSL}}(\varrho_k^{-1} \circ \xi) \\ &= \left[\sum_{k=1}^{K} \widetilde{\alpha}_k \delta(\xi; \xi_{\boldsymbol{s}_k}) \right] * h_{\mathrm{SSL}}(\xi). \end{aligned} \tag{5.59}$$

Here, we absorbed α_k and additional (complex) scaling due to different distances into $\widetilde{\alpha}_k$, and h_{SSL} is computed at some predefined *mean* distance.

We thus reduced the sound source localization problem to a problem of finding the parameters of a weighted sum of Diracs. In order to apply our spherical FRI algorithm, we need to verify that g is bandlimited on the sphere. Figures 5.7D and 5.7E show that it is indeed approximately bandlimited, and that the bandwidth depends on the frequency (it also depends on the sphere radius).

Figures 5.7C and 5.7D show an example of recovering two sources at 1000 Hz and 5 sources at 4000 Hz using the proposed spherical sparse sampling scheme. It is worth noting that this succeeds in spite of the model mismatch due to varying source distances. This indicates the robustness of the proposed reconstruction algorithm.

5.5 Conclusion

We developed a new sampling theorem for sparse signals on the sphere. In particular, by leveraging ideas from finite rate-of-innovation sampling, we showed how to reconstruct sparse collections of spikes on the sphere from their lowpass-filtered observations. Compared to the existing sparse sampling schemes on the sphere, our algorithm uses the available spectrum more efficiently by generalizing known results on 2D harmonic retrieval, thereby reducing the number of samples required to reconstruct the parameters of the spikes.

We illustrated the usefulness of the proposed algorithm by solving three problems: sampling diffusion processes, shot noise removal, and sound source localization. But there is a wealth of other applications, for example in astronomy. Just think about the numerous spherical signal processing challenges put forward by the square kilometer array (SKA) project [49].

We mentioned some approaches to estimation from noisy samples, but more efficient denoising schemes should be studied. One example, effective in the Euclidean setting, is the Cadzow denoising algorithm [29]. The problem seems more challenging on the sphere; in particular, the annihilating matrix is block-Hankel, rather than Hankel.

5.A Annihilating Property

For the sake of completeness, we show in this appendix that the annihilation filter annihilates linear combinations of exponentials. We compute the response of the filter $H(z)$ in (5.33) to a signal of the form $y_n = \sum_{k=1}^{K} b_k x_k^n$ as

$$\begin{aligned}(y * h)_n &= \sum_{m=0}^{K} y_{n-m} h_m = \sum_{m=0}^{K} \left(\sum_{k=1}^{K} b_k x_k^{n-m} \right) h_m \\ &= \sum_{k=1}^{K} x_k^n b_k \sum_{m=0}^{K} h_m x_k^{-m} = \sum_{k=1}^{K} x_k^n b_k \prod_{i=1}^{K} (1 - x_k x_i^{-1}) \\ &= 0.\end{aligned}$$

5.B Computation of the Cramér-Rao Lower Bound

A lowpass-filtered collection of K Diracs can be written as follows,

$$f(\theta, \phi) = \sum_{\ell=0}^{L-1} \sum_{m=-\ell}^{\ell} \left(\sum_{k=1}^{K} \alpha_k \overline{Y_\ell^m(\theta_k, \phi_k)} \right) Y_\ell^m(\theta, \phi). \tag{5.60}$$

In the remainder of this section, we assume $K = 1$, so we rewrite the function as

$$f(\theta, \phi) = \sum_{\ell=0}^{L-1} \sum_{m=-\ell}^{\ell} \alpha_0 \overline{Y_\ell^m(\theta_0, \phi_0)} Y_\ell^m(\theta, \phi). \tag{5.61}$$

We take samples on the sphere at the locations $\{(\theta_n, \phi_n)\}_{n=1}^{N}$. The nth sample is given by

$$\mu_n = f(\theta_n, \phi_n) + \varepsilon_n \tag{5.62}$$

where $\varepsilon_n \sim \mathcal{N}(0, \sigma^2)$, and they are iid. By $\boldsymbol{\zeta} = [\alpha_0,\ \theta_0,\ \phi_0]^\top$, we denote the vector of parameters we estimate. To make the dependence on $\boldsymbol{\zeta}$ explicit, we rewrite (5.62) slightly as

$$\mu_n = f_n(\boldsymbol{\zeta}) + \varepsilon_n. \tag{5.63}$$

With this notation in hand, we can write the conditional probability density function of the nth measurement as

$$p(\mu | \boldsymbol{\zeta}) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-[\mu - f_n(\boldsymbol{\zeta})]^2/(2\sigma^2)}, \tag{5.64}$$

so that the log-likelihood function is

$$\begin{aligned}L(\boldsymbol{\zeta}) &\overset{\text{def}}{=} \ln p(\mu_1, \ldots, \mu_N | \boldsymbol{\zeta}) \\ &= \sum_{n=1}^{N} \left[-\tfrac{1}{2} \ln(2\pi\sigma^2) - (\mu_n - f_n(\boldsymbol{\zeta}))^2/(2\sigma^2) \right].\end{aligned} \tag{5.65}$$

Consequently, differentiating L with respect to any entry ζ_i of $\boldsymbol{\zeta}$ gives

$$\frac{\partial L}{\partial \zeta_i} = \frac{1}{\sigma^2} \sum_{n=1}^{N} \varepsilon_n \frac{\partial f_n(\boldsymbol{\zeta})}{\partial \zeta_i}. \tag{5.66}$$

We can compute the three required derivatives

$$
\begin{aligned}
\frac{\partial f_n(\boldsymbol{\zeta})}{\partial \alpha_0} &= \sum_{\ell=0}^{L-1} \sum_{|m|\leqslant \ell} \overline{Y_\ell^m(\theta_0,\phi_0)} Y_\ell^m(\theta_n,\phi_n), \\
\frac{\partial f_n(\boldsymbol{\zeta})}{\partial \theta_0} &= \alpha_0 \sum_{\ell=0}^{L-1} \sum_{|m|\leqslant \ell} \Bigg[m \cot\theta_0 \overline{Y_\ell^m(\theta_0,\phi_0)} \\
&+ \sqrt{(l-m)(l+m+1)} e^{\mathrm{j}\phi_0} \overline{Y_\ell^{m+1}(\theta_0,\phi_0)} \Bigg] Y_\ell^m(\theta_n,\phi_n), \\
\frac{\partial f_n(\boldsymbol{\zeta})}{\partial \phi_0} &= \alpha_0 \sum_{\ell=0}^{L-1} \sum_{|m|\leqslant \ell} (-\mathrm{j}m) \overline{Y_\ell^m(\theta_0,\phi_0)} Y_\ell^m(\theta_n,\phi_n).
\end{aligned}
$$

Now $\nabla L = \left[\frac{\partial L}{\partial \alpha_0}, \frac{\partial L}{\partial \theta_0}, \frac{\partial L}{\partial \phi_0}\right]^\top$, and the Fisher information matrix is

$$
\boldsymbol{I}(\boldsymbol{\zeta}) \stackrel{\text{def}}{=} \mathbb{E}\left[\nabla L(\boldsymbol{\zeta}) \nabla L(\boldsymbol{\zeta})^H\right] = \frac{1}{\sigma^2} \sum_{n=1}^{N} \nabla f_n(\boldsymbol{\zeta}) \nabla f_n(\boldsymbol{\zeta})^H.
$$

Let $\widehat{\boldsymbol{\zeta}}$ be any unbiased estimator of the parameters $\boldsymbol{\zeta}$. The CRLB can then be computed as

$$
\operatorname{cov}(\widehat{\boldsymbol{\zeta}}) \succeq \boldsymbol{I}(\boldsymbol{\zeta})^{-1}. \tag{5.67}
$$

5.C Rank of the annihilating matrix

In this appendix, we show that the rank of the annihilating matrix $\boldsymbol{Z}$ (5.35) is K with probability one, as soon as it has at least K rows. It then follows follows that the annihilating filter $\boldsymbol{h}$ is uniquely determined, up to a scaling factor, by solving $\boldsymbol{Z}\boldsymbol{h} = \boldsymbol{0}$.

Consider the factorization $\boldsymbol{\Delta} = \boldsymbol{X}\boldsymbol{A}\boldsymbol{U}$,

$$
\boldsymbol{\Delta} = \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_k \\ \vdots & & \vdots \\ x_1^{L-1} & \cdots & x_K^{L-1} \end{bmatrix} \begin{bmatrix} \alpha_1 & & \\ & \ddots & \\ & & \alpha_K \end{bmatrix} \begin{bmatrix} u_{1,1-L} \cdots 1 \cdots u_{1,L-1} \\ \vdots \qquad\qquad \vdots \\ u_{K,1-L} \cdots 1 \cdots u_{1,L-1} \end{bmatrix},
$$

where $x_k = \cos\theta_k$ and $u_{k,m} = (\sin\theta_k)^{|m|} e^{-\mathrm{j}m\phi_k}$.

To construct the annihilating matrix $\boldsymbol{Z}$ as in equation (5.35), we create Hankel blocks from columns of $\boldsymbol{\Delta}$. The $(L-K)\times(K+1)$ Hankel block corresponding to the middle ($m=0$) column of $\boldsymbol{\Delta}$ can be factored as

$$
\begin{aligned}
\boldsymbol{B}_0 &= \begin{bmatrix} x_1^{L-K-1} & \cdots & x_K^{L-K-1} \\ \vdots & & \vdots \\ x_1 & \cdots & x_K \\ 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 & & \\ & \ddots & \\ & & \alpha_K \end{bmatrix} \begin{bmatrix} x_1^K & \cdots & x_1^0 \\ \vdots & & \vdots \\ x_K^K & \cdots & x_K^0 \end{bmatrix} \\
&\stackrel{\text{def}}{=} \boldsymbol{X}_0 \boldsymbol{A} \boldsymbol{\Xi}.
\end{aligned} \tag{5.68}
$$

The second block of the annihilating matrix obtained from the column corresponding to $m = -1$ is similar,

$$\boldsymbol{B}_{-1} = \boldsymbol{X}_1 \boldsymbol{Y}_{-1} \boldsymbol{A}\boldsymbol{\Xi}, \tag{5.69}$$

where $\boldsymbol{Y}_{-1} \overset{\text{def}}{=} \operatorname{diag}(u_{1,-1}, \ldots, u_{K,-1})$, and $\boldsymbol{X}_m$ is obtained by removing the m leading rows from $\boldsymbol{X}_0$. Then we can write

$$\begin{aligned} \boldsymbol{Z} &= [\boldsymbol{B}_0^\top,\ \boldsymbol{B}_{-1}^\top,\ \boldsymbol{B}_1^\top,\ \ldots,\ \boldsymbol{B}_{K-L+1}^\top,\ \boldsymbol{B}_{L-K-1}^\top]^\top \\ &= \begin{bmatrix} \boldsymbol{X}_0 & \cdot\, \boldsymbol{I} & \cdot\, \boldsymbol{A}\boldsymbol{\Xi} \\ \boldsymbol{X}_1 & \cdot\, \boldsymbol{Y}_{-1} & \cdot\, \boldsymbol{A}\boldsymbol{\Xi} \\ \boldsymbol{X}_1 & \cdot\, \boldsymbol{Y}_1 & \cdot\, \boldsymbol{A}\boldsymbol{\Xi} \\ & \vdots & \\ \boldsymbol{X}_{L-K-1} & \cdot\, \boldsymbol{Y}_{K-L+1} & \cdot\, \boldsymbol{A}\boldsymbol{\Xi} \\ \boldsymbol{X}_{L-K-1} & \cdot\, \boldsymbol{Y}_{L-K-1} & \cdot\, \boldsymbol{A}\boldsymbol{\Xi} \end{bmatrix}, \end{aligned} \tag{5.70}$$

with the $\boldsymbol{A}\boldsymbol{\Xi}$ factor being common for all row-blocks. We want to show that the nullspace of $\boldsymbol{Z}$ has dimension one. To that end, we just need to establish that the following matrix,

$$\boldsymbol{T} = \begin{bmatrix} \boldsymbol{X}_0 & \cdot\, \boldsymbol{I} \\ \boldsymbol{X}_1 & \cdot\, \boldsymbol{Y}_{-1} \\ \boldsymbol{X}_1 & \cdot\, \boldsymbol{Y}_1 \\ & \vdots \\ \boldsymbol{X}_{L-K-1} & \cdot\, \boldsymbol{Y}_{K-L+1} \\ \boldsymbol{X}_{L-K-1} & \cdot\, \boldsymbol{Y}_{L-K-1} \end{bmatrix}, \tag{5.71}$$

has full column rank. To see why this is the case, let $\boldsymbol{v}$ be a non-zero vector such that $\boldsymbol{0} = \boldsymbol{Z}\boldsymbol{v} = \boldsymbol{T}(\boldsymbol{A}\boldsymbol{\Xi}\boldsymbol{v})$. It then follows from the full-rankness of $\boldsymbol{T}$ that $\boldsymbol{A}\boldsymbol{\Xi}\boldsymbol{v} = \boldsymbol{0}$. Since $\boldsymbol{A}$ is a diagonal matrix with non-zero entries on the diagonal and $\boldsymbol{\Xi}$ is a $K \times (K+1)$ Vandermonde matrix with distinct roots, the vector $\boldsymbol{v}$ is uniquely determined up to a multiplicative factor. We now show that the matrix $\boldsymbol{T}$ indeed has full column rank almost surely.

Any column in $\boldsymbol{T}^\top$ is of the form

$$\begin{bmatrix} (\cos\theta_1)^r (\sin\theta_1)^{|s|} e^{jj\phi_1 s} \\ \vdots \\ (\cos\theta_K)^r (\sin\theta_K)^{|s|} e^{jj\phi_K s} \end{bmatrix}, \tag{5.72}$$

where $0 \leqslant r < L - K - |s|$ and $-(L-K-1) \leqslant s \leqslant L-K-1$. If the locations of the Diracs are random, we can use the following lemma to show that the matrix $\boldsymbol{T}$ will have full column rank with probability one.

Lemma 5.3

Draw $[\xi_k = (\theta_k, \phi_k)]_{k=1}^K$ independently at random from any absolutely continuous probability distribution on $\mathcal{R} = [0,\pi]\times[0,2\pi]$ (w.r.t. Lebesgue measure). Let $\mathcal{M} = \{(r_1, s_1), \ldots, (r_N, s_N)\}$ be a set of distinct integer pairs and let $\boldsymbol{G} = [g_{pq}]$, where $g_{pq} = (\cos\theta_p)^{r_q}(\sin\theta_p)^{|s_q|} e^{jj\phi_p s_q}$. Then $\boldsymbol{G}$ has full rank almost surely.

Proof: This proof is parallel to that of Theorem 3.2 from [11]. Let $\boldsymbol{G}_M$ be the upper left $M \times M$ minor of $\boldsymbol{G}$. We define the *bad* set $\mathcal{B}_M$ as the set on which $\boldsymbol{G}_M$ is singular,

$$\mathcal{B}_M = \left\{(\xi_1, \ldots, \xi_M) \in \mathcal{R}^M \mid \det \boldsymbol{G}_M = 0\right\}. \tag{5.73}$$

The goal is to show that $\mu(\mathcal{B}_K) = 0$, where μ is the Lebesgue measure on $\mathcal{R}^K$. We proceed by induction on M; for $M = 1$, we have that

$$\boldsymbol{G}_1 = [(\cos\theta_1)^{r_1}(\sin\theta_1)^{|s_1|}\mathrm{e}^{\mathrm{j}\phi_1 s_1}],$$

which is non-zero almost surely, so the claim holds. Assume now that $M < \min(K, N)$ and that the bad set $\mathcal{B}_M$ has measure zero. Let $(\xi_1, \dots, \xi_M) \notin \mathcal{B}_M$, *i.e.*, $\boldsymbol{G}_M$ is invertible. Because it is invertible, there exists a unique coefficient vector $\boldsymbol{b} = \boldsymbol{b}(\xi_1, \dots, \xi_M)$ such that

$$\boldsymbol{G}_M \boldsymbol{b} = \boldsymbol{g}_{M+1}, \tag{5.74}$$

where by $\boldsymbol{g}_{M+1}$ we denote the first M entries of the last column of $\boldsymbol{G}_{M+1}$. The bigger matrix $\boldsymbol{G}_{M+1}$ will be singular if and only if *the same* linear combination is also consistent with its $(M+1)$st row. In other words, $\boldsymbol{G}_{M+1}$ is invertible if and only if ξ_{M+1} is not in the set

$$\mathcal{Z}_M(\xi_1, \dots, \xi_M) = \left\{(\theta, \phi) = \xi \in \mathcal{R} \,\middle|\, (\cos\theta)^{r_{M+1}}(\sin\theta)^{|s_{M+1}|}\mathrm{e}^{\mathrm{j}\phi s_{M+1}} = \sum_{i=1}^{M} b_i(\cos\theta)^{r_i}(\sin\theta)^{|s_i|}\mathrm{e}^{\mathrm{j}\phi s_i}\right\}.$$

For fixed $(\xi_1, \dots, \xi_M)$, this is the set of zeros of a particular (generalized) trigonometric polynomial, thus it has measure zero. Note that the definition of $\mathcal{Z}_M$ makes sense only for $(\xi_1, \dots, \xi_M) \notin \mathcal{B}_M$, as otherwise $\boldsymbol{G}_M$ is not invertible. Thus, the solution $\boldsymbol{b}$ to (5.74) may not exist.

Consider now the following two sets:

$$\mathcal{U}_{M+1} \stackrel{\text{def}}{=} \{(\xi_1, \dots, \xi_{M+1}) \mid (\xi_1, \dots, \xi_M) \in \mathcal{B}_M, \xi_{M+1} \in \mathcal{R}\}$$

and

$$\mathcal{V}_{M+1} \stackrel{\text{def}}{=} \left\{(\xi_1, \dots, \xi_{M+1}) \mid (\xi_1, \dots, \xi_M) \in \mathcal{R}^M, \xi_{M+1} \in \mathcal{Z}_M\right\}.$$

The bad set $\mathcal{B}_{M+1}$ must be a subset of the set $\mathcal{U} \cup \mathcal{V}$. But we just showed that the set $\mathcal{V}$ has measure zero; by the induction hypothesis, $\mathcal{U}$ also has measure zero. Thus their union, too, has measure zero.

It follows that $\mathcal{B}_{M+1}$ has measure zero. Finally, because the distributions of ξ_i are absolutely continuous w.r.t. the Lebesgue measure, so is their product distribution. Hence the probability that $(\xi_1, \dots, \xi_K)$ lies in the zero-measure set $\mathcal{B}_K$ is zero. ■

To complete the argument, note that the matrix $\boldsymbol{T}^\top$ has the same form as the matrix $\boldsymbol{G}$ in the statement of Lemma 5.3, with $0 \leqslant r < L - K - |s|$ and $-(L - K - 1) \leqslant s \leqslant L - K - 1$. Thus, the columns of $\boldsymbol{T}$ are independent with probability one, provided that its number of rows is at least K.

Chapter 6

Acoustic Rake Receiver[*]

6.1 Introduction

In this chapter we propose acoustic rake receivers[1] (ARR)—microphone beamformers that use echoes to improve noise and interference suppression. Instead of trying to mitigate multipath propagation, rake receivers take advantage of it. The basic idea of the rake receivers, which are commonly used in wireless communications, is to coherently add the multipath components and thus increase the effective signal-to-noise ratio (SNR).

Rake receivers were introduced by Price and Green [171], who described a wideband radio receiver consisting of a delay line and a series of correlators synchronized at successive delay increments of $1/W$, where W is the signal bandwidth. Under certain assumptions, this system is capable of discriminating and "bringing into step" all multipath components, which are then summed algebraically instead of vectorially. The bank of correlators synchronized at different delays resembles the fingers of a rake[2] that collects, or *rakes*, the various signal paths.

The original scheme [171] was developed for single-antenna systems, and it was later extended to arrays of antennas by Khalaj, Paulraj, and Kailath [106] and by Naguib [155]. The advantage of using an antenna array is that we get spatial selectivity. Thus components that are not resolvable with a single antenna because they arrive at similar times become resolvable because they arrive from different directions. One should imagine a beamformer whose beam is steered towards each of the components in turn.

6.1.1 Wireless vs Speech

In spite of the success of the rake receivers in wireless communications, the principle has not received significant attention in room acoustics. Nevertheless, constructive use of echoes in rooms to improve beamforming has been mentioned in the literature [6, 160, 96]. In particular, the term *Acoustic Rake Receiver* (ARR) was used in the SCENIC project proposal [6].

---

[1] We caution the reader that ARR is somewhat of a misnomer as we primarily exploit the spatial structure of the multipath, unlike the original work on rake receivers (and the origin of the noun adjunct *rake*) that was exploiting the temporal structure.

[2] A rake is a broom for outside use; a horticultural implement consisting of a toothed bar fixed transversely to a handle, and used to collect leaves, hay, grass, etc. [Wikipedia].

[*] This chapter is a result of a joint work with Robin Scheibler and Martin Vetterli[60, 184].

The list of ingredients for ARRs in room acoustics is similar as in wireless communications: a wave (acoustic instead of electromagnetic) propagates in space; reflections and scattering cause the wave to arrive at the receiver through multiple paths in addition to the direct path, and these multipath components all contain the source waveform.

The main difference is that in room acoustics we do not get to design the input signal. What makes rake receivers work in wireless communications is the particular signal structure. Price and Green [171] assume that the sender sends two different waveforms, “Mark” and “Space”, which have a low cross-correlation and a very peaky autocorrelation. Khalaj, Paulraj and Kailath [106] use known pseudo-noise spreading (chip) sequences with similar properties (near-orthogonality to shifts, and orthogonality between different users), as is common in spread spectrum CDMA systems [169, 181].

All this facilitates multipath channel estimation: The basic approach is simply to crosscorrelate the received signals with known templates. But with speech we have no such structure. Even if the inverse bandwidth of speech is relatively small compared to typical separations between strong echoes, speech segments are long and unknown. We have an idea about their statistics, but there is no firm *template* that we could use to estimate the multipath structure.

On the contrary, there are no significant differences between wireless communications and speech in terms of the spatial structure. If we know where the echoes are coming from, we can design spatial processing algorithms—for example beamformers—that use multiple copies of the same signal arriving from different directions. The most significant difference is perhaps that the resulting signal is to be listened to by people, leading to additional perceptual constraints on processing.

Imagine first that we know the room geometry. Then, if we localize the source, we can predict where its echoes will come from by using simple geometric rules [5, 19]. Localizing the direct signal in a reverberant environment is a well-understood problem [218, 50]. What is more, we do not need to know the room shape in detail—locations of the most important reflectors (ceiling, floor, walls) suffice to localize the major echoes. In many cases this knowledge is readily available from floor plans or measurements. In ad-hoc deployments, the room geometry may be difficult to obtain. If that is the case, we can first perform a calibration step to learn it. An appealing method to infer the room geometry is by using sound and the same array we use for beamforming, as was demonstrated recently [176, 7, 54, 56].

We may still be able to take advantage of the echoes without estimating the room geometry. Note that we are not after the room geometry itself; rather, we only need to know where the early echoes are coming from. Echoes can be seen as signals emitted by *image sources*—mirror images of the true source across reflecting walls [5]. Knowing where the echoes are coming from is equivalent to knowing where the image sources are.

Image source localization can be solved, for example, by *echo sorting* as described in [56]. Alternatively, O’Donovan, Duraiswami, and Zotkin [160] propose to use an *audio camera* with a large number of microphones to find the images. Once the image sources are localized (in a calibration phase or otherwise), we can predict their movement using geometrical rules, as discussed in Section 6.5. Thus, the acoustic raking is a multi-stage process comprising real and image source localization and tracking, and the computation of beamforming weights. The complete block diagram is shown in Figure 6.1.

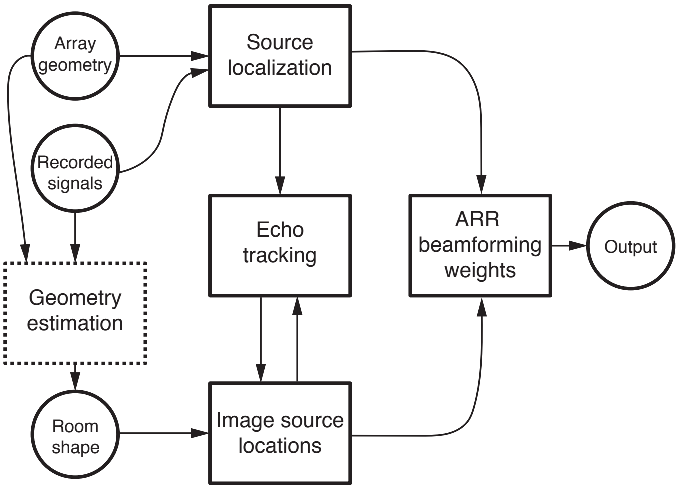


Figure 6.1: A block diagram for acoustic rake receivers. In this chapter, we focus on ARR beamforming weight computation, and we briefly discuss echo tracking and image source localization. The geometry estimation block is optional (room geometry could be known in advance), hence the dashed box.

6.1.2 Related Work

It is interesting to note the analogy between the ARRs and human auditory perception. It is well established that the early echoes improve speech intelligibility [25, 127]. In fact, adding energy in the form of early echoes—approximately within the first 50 ms of the room impulse response (RIR)—is equivalent to adding the same energy to the direct sound [25]. This observation suggests new designs for indoor beamformers, with different choices of performance measures and reference signals. A related discussion of this topic is given by Habets and co-authors [88], who examine the tradeoff between dereverberation and denoising in beamforming. In addition to the standard SNR, we propose to use the useful-to-detrimental ratio (UDR), first defined by Lochner and Burger [127], and used by Bradley, Sato and Picard [25]. We generalize UDR to a scenario with interferers, defining it as the the ratio of the direct and early reflection energy to the energy of the noise and interference.

ARRs focus on the early part of the RIR, trying to concentrate the energy contained in the early echoes. In that regard, there are similarities between ARRs and channel shortening [201, 223]. Channel shortening produces filters that are much better behaved than complete inversion, *e.g.*, by the multiple-input-output-theorem (MINT) [148, 75]. Nevertheless, it is assumed that we know the acoustic impulse responses between the sources and the microphones. In contrast to channel shortening, as well as other methods assuming this knowledge [16, 148], we never attempt the difficult task of estimating the impulse responses. Our task is simpler: we only need to detect the early echoes, and lift them to 3D space as image sources.

6.1.3 Main Contributions and Limitations

We introduce the acoustic rake receiver (ARR) as an echo-aware microphone beamformer. We present several formulations with different properties, and analyze their behavior theoretically and numerically. The analysis shows that ARRs lead to significantly improved SNR and interference cancellation when compared with standard beamformers that only extract the direct path. ARRs can suppress interference in cases when conventional beamforming is bound to fail, for example when an interferer is occluding the desired source. This is illustrated in Figure 6.2 (for a sneak-peak of the numerical results, fast forward to Figure 6.8). We can *listen behind* an interferer by listening to echoes of the desired source, instead of listening to its direct path. This could be seen as an analogy with user separation in multiuser wireless systems; however, the latter is achieved primarily by temporal means (orthogonal codes).

The raking microphone beamformers are particularly well-suited to extracting the desired speech signal in the presence of interfering sounds, in part because they can focus on echoes of the desired sound and cancel the echoes of the interference. The analogous human capacity to focus on a particular acoustic stimulus while not perceiving other, unwanted sounds is called the *cocktail party effect* [91].

We present optimal formulations that outperform the earlier delay-and-sum (DS) approaches [96], especially when interferers are present. Significant gains are observed not only in terms of signal-to-interference-and-noise ratio (SINR) and UDR, but also in terms of perceptual evaluation of speech quality (PESQ) [178].

We first design and apply the ARRs in the frequency domain. Frequency domain formulation is simple and concise; it allows us to focus on objective gains from acoustic raking; time-domain designs [92, 183] offer better control over the impulse responses of the beamforming filters. Towards the end of the chapter we propose the time-domain formulation of the rake receivers, and first experimental results that show that these receivers indeed improve the beamforming impulse responses.

Let us also mention some limitations of our results. For clarity, the numerical experiments are presented in a 2D "room", and as such are directly applicable to planar (*e.g.*, linear or circular) arrays. Extension to 3D arrays is straightforward. We do not discuss robust formulations that address uncertainties in the array calibration. Microphones are assumed to be ideally omnidirectional with a flat frequency response. Except for Section 6.5, we assume that the locations of the image sources are known. We explain how to find the image sources when the room geometry is either known or unknown; for details about room geometry estimation techniques, as well as the list of references, we refer the reader to Chapter 3. We consider the walls to be flat-fading; in reality, they are frequency selective. We do not discuss the estimation of various covariance matrices [32].

6.2 Signal model

Suppose that the desired source of sound is at the location $\boldsymbol{s}_0$ in a room. Sound from this source arrives at the microphones located at $\{\boldsymbol{r}_m\}_{m=1}^M$ via the direct path, but also via the echoes from the walls. The echoes can be replaced by the image sources—mirror images of the true sources across the corresponding walls—according to the image source model (*cf.* Chapter 2). An important consequence is that instead of modeling the source of the desired or the interefering signal as a single point in a room, we can model it as a collection of points in free space.

Denote by $x[n]$ the signal[3] (*e.g.* speech) emitted by the source; then all the image sources emit

---

[3] We use the same symbol (*e.g.* x) for both the time-domain signal $x[n]$ and its spectrum $x(e^{j\omega})$. The reason for this choice is to keep the *clean* symbol for the frequency domain equations and avoid clutter, while not introducing a non-standard symbol for the time-domain signals. If the argument is absent, we assume the frequency domain.

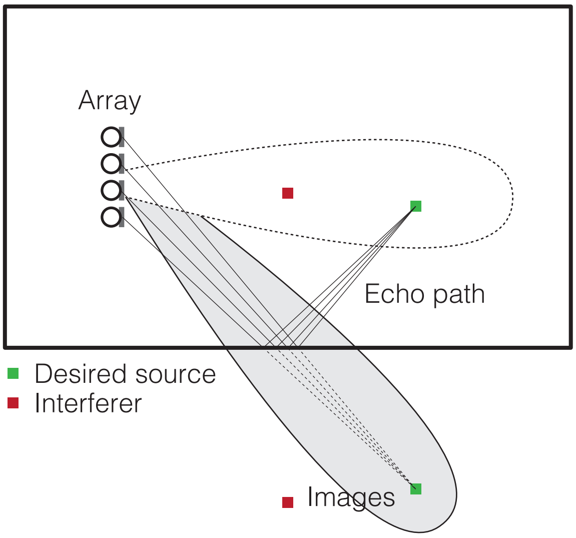


Figure 6.2: Listening behind an interferer by listening to echoes (illustration).

$x[n]$ as well, and the signals from the image sources reach the microphones with the appropriate delays. In our application, the essential fact is that echoes correspond to image sources. We denote the image source positions by $\boldsymbol{s}_k$, $1 \leqslant k \leqslant K$, where K denotes the largest number of image sources considered. Note that we do not care about the sequence of walls that generates $\boldsymbol{s}_k$, nor do we care about how many walls are in this sequence. For us, all $\boldsymbol{s}_k$ are simply additional sources of the desired signal. The described setup is illustrated Figure 6.3.

Suppose further that there is an interferer at the location $\boldsymbol{q}_0$ (for simplicity, we consider only a single interferer). The interferer emits a signal $z[n]$, and its image sources emit $z[n]$ as well. Similarly as for the desired source, we denote by $\boldsymbol{q}_k$, $1 \leqslant k \leqslant K'$ the positions of the interfering image sources, with K' being the largest number of interfering image source considered. The model mismatch (*e.g.*, the image sources of high orders and the late reverberation) and the noise are absorbed in the term $n_m[n]$.

The signal received by the mth microphone is then a sum of convolutions

$$y_m[n] = \sum_{k=0}^{K} \big(a_m(\boldsymbol{s}_k) * x\big)[n] + \sum_{k=0}^{K'} \big(a_m(\boldsymbol{q}_k) * z\big)[n] + n_m[n], \tag{6.1}$$

where $a_m(\boldsymbol{s}_k)$ denotes the impulse response of the channel between the source located at $\boldsymbol{s}_k$ and the mth microphone—in this case a delay and a scaling factor.

We will first discuss beamforming in the frequency domain, as it is conceptually simpler. That is, we will be working with the DTFT of the discrete-time signal x,

$$x(\mathrm{e}^{\mathrm{j}\omega}) \overset{\text{def}}{=} \sum_{n\in\mathbb{Z}} x[n]\, \mathrm{e}^{-\mathrm{j}\omega n}. \tag{6.2}$$

In practical implementations, we use the discrete-time short-time Fourier transform (STFT). More implementation details are given in Section 6.6.

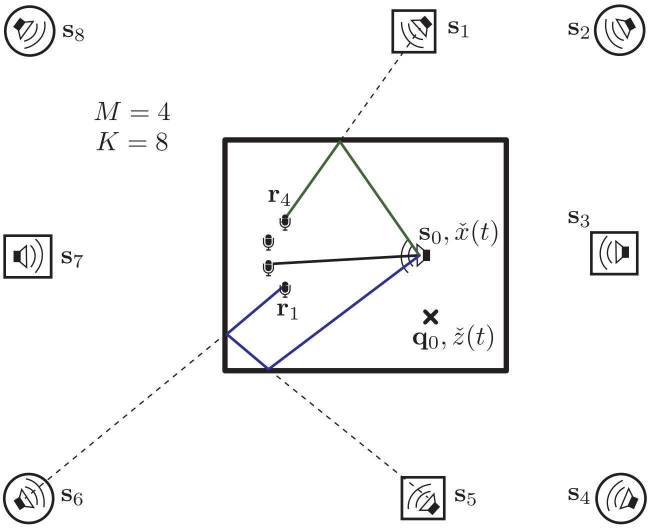


Figure 6.3: Illustration of the notation and concepts. Echoes of the desired signal emitted at $\boldsymbol{s}_0$ can be modeled as a direct sound coming from the image sources of $\boldsymbol{s}_0$. Two generations of image sources are illustrated: first ($\boldsymbol{s}_1, \boldsymbol{s}_3, \boldsymbol{s}_5, \boldsymbol{s}_7$) and second ($\boldsymbol{s}_2, \boldsymbol{s}_4, \boldsymbol{s}_6, \boldsymbol{s}_8$), as well as the corresponding *sound rays* for $\boldsymbol{s}_5$ and $\boldsymbol{s}_6$. The interferer is located at $\boldsymbol{q}_0$ (its image sources are not shown), and the microphones are located at $\boldsymbol{r}_1, \ldots, \boldsymbol{r}_4$.

Using these notations, we can write the signal picked up by the mth microphone as

$$y_m(\mathrm{e}^{\mathrm{j}\omega}) = \sum_{k=0}^{K} a_m(\boldsymbol{s}_k, \Omega) x(\mathrm{e}^{\mathrm{j}\omega}) + \sum_{k=0}^{K'} a_m(\boldsymbol{q}_k, \Omega) z(\mathrm{e}^{\mathrm{j}\omega}) + n_m(\mathrm{e}^{\mathrm{j}\omega}), \tag{6.3}$$

where $n_m(\mathrm{e}^{\mathrm{j}\omega})$ models the noise and other errors, and $a_m(\boldsymbol{s}_k, \Omega)$ denotes the mth component of the steering vector for the source $\boldsymbol{s}_k$. The steering vector is the Fourier transform of the continuous version of the impulse response $a(\boldsymbol{s}_k)$, evaluated at the frequency Ω. The discrete-time frequency ω and the continuous-time frequency Ω are related as $\omega = \Omega T_s$, where T_s is the sampling period. The steering vector is then simply $\boldsymbol{a}(\boldsymbol{s}_k, \Omega) = [a_m(\boldsymbol{s}_k, \Omega)]_{m=0}^{M-1}$.

We can write out the entries of the steering vectors explicitly for a point source in free space. They are given as the appropriately scaled free-space Green's functions for the Helmholtz equation [67] (*cf.* Chapter 2, equation (2.58)),

$$a_m(\boldsymbol{s}_k, \Omega) = \frac{\alpha_k}{4\pi \|\boldsymbol{r}_m - \boldsymbol{s}_k\|} \mathrm{e}^{-\mathrm{j}\kappa \|\boldsymbol{r}_m - \boldsymbol{s}_k\|}, \tag{6.4}$$

where we define the wavenumber as $\kappa \overset{\text{def}}{=} \Omega / c$, and α_k is the attenuation corresponding to $\boldsymbol{s}_k$.

Using vector notation, the microphone signals can be written concisely as

$$\boldsymbol{y}(\mathrm{e}^{\mathrm{j}\omega}) = \boldsymbol{A}_s(\mathrm{e}^{\mathrm{j}\omega}) \boldsymbol{1} x(\mathrm{e}^{\mathrm{j}\omega}) + \boldsymbol{A}_q(\mathrm{e}^{\mathrm{j}\omega}) \boldsymbol{1} z(\mathrm{e}^{\mathrm{j}\omega}) + \boldsymbol{n}(\mathrm{e}^{\mathrm{j}\omega}), \tag{6.5}$$

where $\boldsymbol{A}_s(\mathrm{e}^{\mathrm{j}\omega}) \overset{\text{def}}{=} [\boldsymbol{a}(\boldsymbol{s}_1, \Omega), \ldots, \boldsymbol{a}(\boldsymbol{s}_K, \Omega)]$, $\boldsymbol{A}_q(\mathrm{e}^{\mathrm{j}\omega}) \overset{\text{def}}{=} [\boldsymbol{a}(\boldsymbol{q}_1, \Omega), \ldots, \boldsymbol{a}(\boldsymbol{q}_{K'}, \Omega)]$, and $\boldsymbol{1}$ is the all-ones vector.

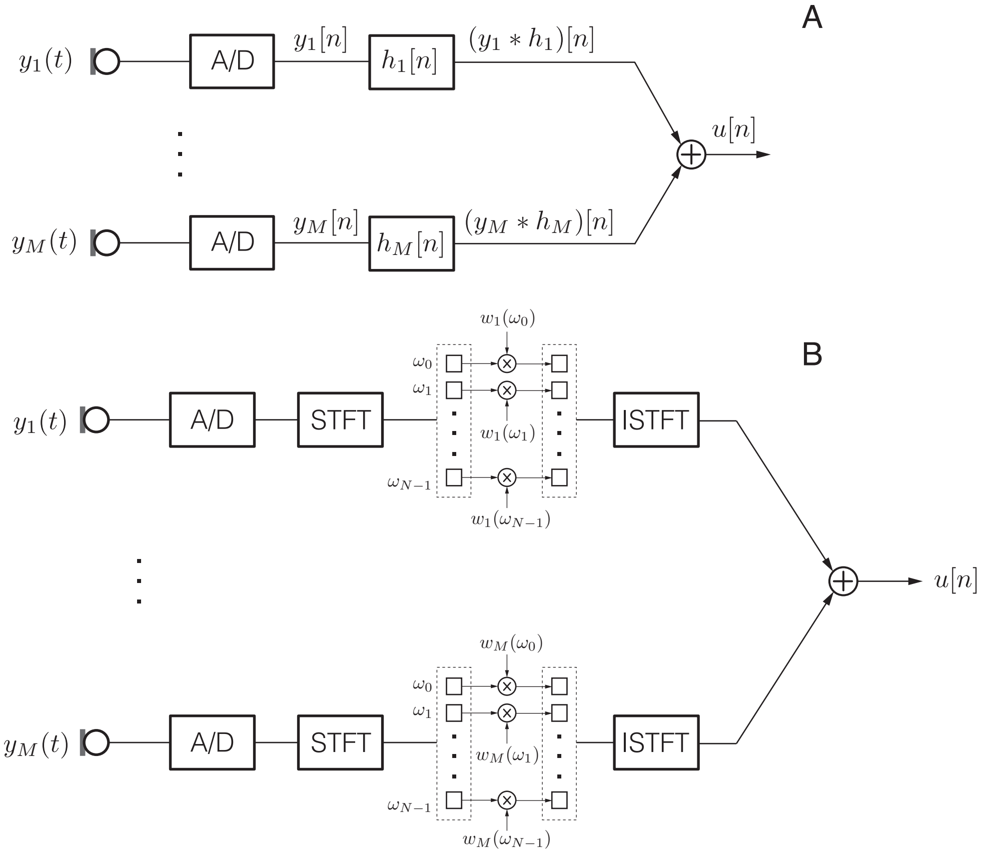


Figure 6.4: Stucture of time-domain and frequency-domain beamformers. (A) In the time domain, signals received by the microphones are filtered by the filters h_m (often FIR). (B) In the frequency domain, microphone signals are typically processed with the short-time Fourier transform (STFT) and then multiplied per-frequency by the corresponding frequency-dependent beamforming weights.

6.3 Beamforming Preliminaries

Microphone beamformers combine the outputs of multiple microphones in order to achieve spatial selectivity, or more generally to suppress noise and interference [210, 199]. Basic structures of frequency-domain and time-domain beamformers are illustrated in Figure 6.4. The design parameters are filters h_m in the time domain, and weights $w_m(\mathrm{e}^{\mathrm{j}\omega})$ in the frequency domain. The goal is to design these filters and weights in a way that will optimize certain design criteria, such as interference suppression or SNR at the output.

We primarily treat beamforming in the frequency domain because it is conceptually simpler (this will be clear later when we discuss time-domain raking). With reference to the figure, we see that the output of the time-domain beamformer is computed as

$$u[n] = \sum_{m=1}^{M} (y_m * h_m)[n]. \tag{6.6}$$

Taking the DTFT of both sides we have that

$$u(\mathrm{e}^{\mathrm{j}\omega}) = \sum_{m=1}^{M} y_m(\mathrm{e}^{\mathrm{j}\omega}) h_m(\mathrm{e}^{\mathrm{j}\omega}). \tag{6.7}$$

Evidently, for a narrowband signal, a beamformer forms a simple linear combination of the microphone outputs to yield the output u. A typical procedure is then to design beamformers independently in every frequency bin (although there are exceptions to this, for example when we desire constant beam shape over a wide frequency band [219, 79]).

From here onward, we suppress the frequency dependency of the steering vectors and the beamforming weights to reduce the notational clutter. We can write (6.7) as

$$u = \boldsymbol{w}^* \boldsymbol{y} = \boldsymbol{w}^* \boldsymbol{A}_s \boldsymbol{1} x + \boldsymbol{w}^* \boldsymbol{A}_q \boldsymbol{1} z + \boldsymbol{w}^* \boldsymbol{n}, \tag{6.8}$$

where the vector $\boldsymbol{w} \in \mathbb{C}^M$ contains the beamforming weights.

The weights $\boldsymbol{w}$ are often selected so that they optimize some design criterion. Common examples of beamformers are the delay-and-sum (DS) beamformer, minimum-variance-distortionless-response (MVDR) beamformer, maximum-signal-to-interference-and-noise (Max-SINR) beamformer, and minimum-mean-squared-error (MMSE) beamformer. In this chapter we discuss the rake formulation of the DS and the Max-SINR beamformers; for completeness, we first describe the non-raking variants.

6.3.1 Delay-and-Sum Beamformer

DS is the simplest and often quite effective beamformer [210]. It inserts delays into microphone signals so that they become aligned in time, and then takes their mean. Assume that we want to listen to a source at $\boldsymbol{s}$. Then we form the DS beamformer by compensating the propagation delays from the source $\boldsymbol{s}$ to the microphones $\boldsymbol{r}_m$,

$$\begin{aligned} u_{\mathrm{DS}} = \frac{1}{M} \sum_{m=0}^{M-1} y_m \mathrm{e}^{\mathrm{j}\kappa \|\boldsymbol{r}_m - \boldsymbol{s}\|} &= \frac{1}{M} \sum_{m=0}^{M-1} \left[\frac{x\, \mathrm{e}^{-\mathrm{j}\kappa \|\boldsymbol{r}_m - \boldsymbol{s}\|}}{4\pi \|\boldsymbol{r}_m - \boldsymbol{s}\|} + n_m \right] \mathrm{e}^{\mathrm{j}\kappa \|\boldsymbol{r}_m - \boldsymbol{s}\|} && (6.9) \\ &\approx \frac{x}{4\pi \|\overline{\boldsymbol{r}} - \boldsymbol{s}\|} + \frac{1}{M} \sum_{m=0}^{M-1} \mathrm{e}^{\mathrm{j}\kappa \|\boldsymbol{r}_m - \boldsymbol{s}\|} n_m && (6.10) \\ &= \overline{y} + n, && (6.11) \end{aligned}$$

where $\overline{\boldsymbol{r}} = \frac{1}{M} \sum_{m=0}^{M-1} \boldsymbol{r}_m$ denotes the center of the array. The beamforming weights can be read out from (6.9) as

$$\boldsymbol{w}_{\mathrm{DS}} = \frac{\boldsymbol{a}(\boldsymbol{s})}{\|\boldsymbol{a}(\boldsymbol{s})\|}, \tag{6.12}$$

where we used the definition of y_m (6.5) and the definition of the steering vector (6.4). We can see from (6.9) that if $\boldsymbol{n} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_M)$, then the output noise n is distributed according to $\mathcal{N}(0, \sigma^2/M)$, that is, we obtain an M-fold decrease in the noise variance at the output with respect to any reference microphone.

6.3.2 Maximum Signal-to-Interference-and-Noise Ratio Beamformer

The SINR is an important figure of merit used to assess the performance of ARRs, and to compare it with the standard non-raking beamformers. It is computed as the ratio of the power of the desired output signal to the power of the undesired output signal. The desired output signal is the output signal due to the desired source, while the undesired signal is the output signal due to the interferers and noise.

For a desired source at $\boldsymbol{s}$ and an interfering source at $\boldsymbol{q}$ we can write

$$\mathsf{SINR} \overset{\text{def}}{=} \frac{\mathbb{E}|\boldsymbol{w}^* \boldsymbol{a}(\boldsymbol{s})x|^2}{\mathbb{E}|\boldsymbol{w}^*(\boldsymbol{a}(\boldsymbol{q})z + \boldsymbol{n})|^2} = \sigma_x^2 \frac{\boldsymbol{w}^* \boldsymbol{a}(\boldsymbol{s}) \boldsymbol{a}(\boldsymbol{s})^* \boldsymbol{w}}{\boldsymbol{w}^* \boldsymbol{K}_{nq} \boldsymbol{w}}, \tag{6.13}$$

where $\boldsymbol{K}_{nq}$ is the covariance matrix of the noise and the interference.

It is compelling to pick $\boldsymbol{w}$ that maximizes the SINR (6.13) [210]. The maximization can be solved by noting that the rescaling of the beamformer weights leaves the SINR unchanged. This means that we can minimize the denominator subject to numerator being an arbitrary constant. The solution is given as

$$\boldsymbol{w}_{\text{SINR}} = \frac{\boldsymbol{K}_{nq}^{-1} \boldsymbol{a}_s}{\boldsymbol{a}_s^* \boldsymbol{K}_{nq}^{-1} \boldsymbol{a}_s}. \tag{6.14}$$

Using the definition (6.13), we can derive the SINR for the Max-SINR beamformer as

$$\mathsf{SINR} = \sigma_x^2 \boldsymbol{a}_s^* \boldsymbol{K}_{nq}^{-1} \boldsymbol{a}_s. \tag{6.15}$$

Because $\boldsymbol{K}_{nq}^{-1}$ is a Hermitian symmetric positive definite matrix, it has an eigenvalue decomposition as $\boldsymbol{K}_{nq}^{-1} = \boldsymbol{U}^* \boldsymbol{\Lambda} \boldsymbol{U}$, where $\boldsymbol{U}$ is unitary, and $\boldsymbol{\Lambda}$ is diagonal with positive entries. We can write $\boldsymbol{a}^* \boldsymbol{K}_{nq}^{-1} \boldsymbol{a} = (\boldsymbol{U}\boldsymbol{a})^* \boldsymbol{\Lambda} (\boldsymbol{U}\boldsymbol{a})$. Because $\|\boldsymbol{U}\boldsymbol{a}\|^2 = \|\boldsymbol{a}\|^2$, and because $\boldsymbol{\Lambda}$ is positive, increasing $\|\boldsymbol{a}\|^2$ typically leads to an increased SINR, although we can construct counterexamples. This will be important when we discuss the SINR gain of the Rake-Max-SINR beamformer.

Note that we do not discuss the well-known MVDR beamformer as in the narrowband case it is the same (up to a scaling of the output) as the Max-SINR beamformer when all covariance matrices are accurately known (as we assume).

6.4 Acoustic Rake Receivers

In this section, we present several formulations of the ARR. The Rake-DS beamformer is a straightforward generalization of the conventional DS beamformer. The one-forcing beamformer implements the idea of steering a fixed beam power towards every image source, while trying to minimize interference and noise. It is a naive extension of the MVDR beamformer in that it attempts to get *undistorted* versions of each image sources. In practice it performs poorly, but we include it to show how intuition can lead to very bad results. The Rake-Max-SINR and Rake-Max-UDR beamformers optimize the corresonding performance measures; we show in Section 6.6 that the Rake-Max-SINR beamforming performs best (except, as expected, in terms of UDR).

6.4.1 Delay-and-Sum Raking

If we had access to every echo separately (*i.e.* not summed with all the other echoes), we could align them all to maximize the performance gain. Unfortunately, this is not the case: each microphone picks up the convolution of speech with the impulse response, which is effectively a sum of overlapping echoes of the speech signal. If we only wanted to extract the direct path, we would use the standard DS beamformer (6.12). To build the Rake-DS receiver, we create a DS beamformer for every image source, and average the outputs,

$$\frac{1}{K+1}\sum_{k=0}^{K}\frac{\alpha_k'}{M}\sum_{m=0}^{M-1} y_m \mathrm{e}^{\mathrm{j}\kappa\|\boldsymbol{r}_m-\boldsymbol{s}_k\|}, \tag{6.16}$$

where $\alpha_k' \overset{\text{def}}{=} \alpha_k/(4\pi\|\boldsymbol{r}_m - \boldsymbol{s}_k\|)$. We read out the beamforming weights from (6.16) as

$$\boldsymbol{w}_{\text{R-DS}} = \frac{1}{\|\sum_k \boldsymbol{a}(\boldsymbol{s}_k)\|}\sum_{k=0}^{K}\boldsymbol{a}(\boldsymbol{s}_k) = \frac{\boldsymbol{A}_s\boldsymbol{1}}{\|\boldsymbol{A}_s\boldsymbol{1}\|}, \tag{6.17}$$

where we chose the scaling in analogy with (6.12) (scaling of the weights does not alter the output SINR). Thus the weights for the Rake-DS beamformer are just a scaled sum of the steering vectors for each image source.

6.4.2 One-Forcing Raking

A different approach, based on intuition, is to design a beamformer that listens to all K image sources with the same power, and at the same time minimizes the noise and interference energy:

$$\begin{aligned} \underset{\boldsymbol{w}\in\mathbb{C}^M}{\text{minimize}}\ & \mathbb{E}\left|\sum_{k=0}^{K'}\boldsymbol{w}^*\boldsymbol{a}(\boldsymbol{q}_k)z + \boldsymbol{w}^*\boldsymbol{n}\right|^2 \\ \text{subject to}\ & \boldsymbol{w}^*\boldsymbol{a}(\boldsymbol{s}_k) = 1, \forall\ 0 \leqslant k \leqslant K. \end{aligned} \tag{6.18}$$

Alternatively, we may choose to null the interfering source and its image sources. Both cases are an instance of the standard linearly-constrained-minimum-variance (LCMV) beamformer [74]. Collecting all the steering vectors in a matrix, we can write the constraint as $\boldsymbol{w}^*\boldsymbol{A}_s = \boldsymbol{1}^\top$. The solution can be found in closed form as

$$\boldsymbol{w}_{\text{R-OF}} = \boldsymbol{K}_{nq}^{-1}\boldsymbol{A}_s(\boldsymbol{A}_s^*\boldsymbol{K}_{nq}^{-1}\boldsymbol{A}_s)^{-1}\boldsymbol{1}_M. \tag{6.19}$$

A few remarks are in order. First, with M microphones, it does not make sense to increase K beyond M, as this results in more constraints than degrees of freedom. Second, using this beamformer is a bad idea if there is an interferer along the ray through the microphone array and any of image sources.

As with all LCMV beamformers, adding linear constraints uses up degrees of freedom that could otherwise be used for noise and interference suppression. It is better to let the "beamformer decide" or "the beamforming procedure decide" on how to maximize a well-chosen cost function; one such procedure is described in the next subsection.

6.4.3 Max-SINR Raking

The main workhorse of the paper is the Rake-Max-SINR. We compute the weights so as to maximize the SINR, taking into account the echoes of the desired signal, and the echoes of the interfering signal,

$$\underset{\boldsymbol{w}\in\mathbb{C}^M}{\text{maximize}}\ \frac{\mathbb{E}\left|\sum_{k=0}^{K}\boldsymbol{w}^*\boldsymbol{a}(\boldsymbol{s}_k)x\right|^2}{\mathbb{E}\left|\sum_{k=0}^{K'}\boldsymbol{w}^*\boldsymbol{a}(\boldsymbol{q}_k)z + \boldsymbol{w}^*\boldsymbol{n}\right|^2}. \tag{6.20}$$

Table 6.1: Summary of beamformers.

Acronym	**Description**	**Beamforming Weights**
DS	Align delayed copies of signal at the microphone	$\boldsymbol{w}_{\text{DS}} = \boldsymbol{a}(\boldsymbol{s})/\Vert\boldsymbol{a}(\boldsymbol{s})\Vert$
Max-SINR	max. $\boldsymbol{w}^*\boldsymbol{a}_s\boldsymbol{a}_s^*\boldsymbol{w}/(\boldsymbol{w}^*\boldsymbol{K}_{nq}\boldsymbol{w})$	$\boldsymbol{w}_{\text{SINR}} = \boldsymbol{K}_{nq}^{-1}\boldsymbol{a}_s/(\boldsymbol{a}_s^*\boldsymbol{K}_{nq}^{-1}\boldsymbol{a}_s)$
Rake-DS	Weighted average of DS beamformers over image sources	$\boldsymbol{w}_{\text{R-DS}} = \boldsymbol{A}_s\boldsymbol{1}/\Vert\boldsymbol{A}_s\boldsymbol{1}\Vert$
Rake-OF	min. $\mathbb{E}\left\lvert\sum_{k=0}^{K'}\boldsymbol{w}^*\boldsymbol{a}(\boldsymbol{q}_k)z + \boldsymbol{w}^*\boldsymbol{n}\right\rvert^2$, s.t. $\boldsymbol{w}^*\boldsymbol{A}_s = \boldsymbol{1}^\top$	$\boldsymbol{w}_{\text{R-OF}} = \boldsymbol{K}_{nq}^{-1}\boldsymbol{A}_s(\boldsymbol{A}_s^*\boldsymbol{K}_{nq}^{-1}\boldsymbol{A}_s)^{-1}\boldsymbol{1}_M$
Rake-Max-SINR	max. $\mathbb{E}\left\lvert\sum_{k=0}^{K}\boldsymbol{w}^*\boldsymbol{a}(\boldsymbol{s}_k)x\right\rvert^2 / \ \mathbb{E}\left\lvert\sum_{k=0}^{K'}\boldsymbol{w}^*\boldsymbol{a}(\boldsymbol{q}_k)z + \boldsymbol{w}^*\boldsymbol{n}\right\rvert^2$	$\boldsymbol{w}_{\text{R-SINR}} = \boldsymbol{K}_{nq}^{-1}\boldsymbol{A}_s\boldsymbol{1}/(\boldsymbol{1}^*\boldsymbol{A}_s^*\boldsymbol{K}_{nq}^{-1}\boldsymbol{A}_s\boldsymbol{1})$
Rake-Max-UDR	max. $\mathbb{E}\sum_{k=0}^{K}\lvert\boldsymbol{w}^*\boldsymbol{a}(\boldsymbol{s}_k)x\rvert^2 / \ \mathbb{E}\left\lvert\boldsymbol{w}^*\sum_{k=0}^{K'}\boldsymbol{a}(\boldsymbol{q}_k)z + \boldsymbol{w}^*\boldsymbol{n}\right\rvert^2$	$\boldsymbol{w}_{\text{R-UDR}} = \boldsymbol{w}_{\max},\ \boldsymbol{A}_s\boldsymbol{A}_s^*\boldsymbol{w}_{\max} = \lambda_{\max}\boldsymbol{K}_{nq}\boldsymbol{w}_{\max}$

The logic behind this expression can be summarized as follows: we present the beamforming procedure with a set of good sources whose influence we aim to maximize at the output, and with a set of bad sources whose power we try to minimize at the output. It turns out that this leads to the standard Max-SINR beamformer with a structured steering vector and covariance matrix. We define the combined noise and interference covariance matrix as

$$\boldsymbol{K}_{nq} \stackrel{\text{def}}{=} \boldsymbol{K}_n + \sigma_z^2 \left(\sum_{k=0}^{K'} \boldsymbol{a}(\boldsymbol{q}_k)\right)\left(\sum_{k=0}^{K'} \boldsymbol{a}(\boldsymbol{q}_k)\right)^*, \tag{6.21}$$

where $\boldsymbol{K}_n$ is the covariance matrix of the noise term, and σ_z^2 is the power of the interferer at a particular frequency. Then the solution to (6.20) is given as

$$\boldsymbol{w}_{\text{R-SINR}} = \frac{\boldsymbol{K}_{nq}^{-1}\boldsymbol{A}_s\boldsymbol{1}}{\boldsymbol{1}^*\boldsymbol{A}_s^*\boldsymbol{K}_{nq}^{-1}\boldsymbol{A}_s\boldsymbol{1}}. \tag{6.22}$$

Note that when $\boldsymbol{K}_{nq} = \sigma^2\boldsymbol{I}_M$ (*e.g.* no interferers and iid noise), the Rake-Max-SINR beamformer reduces to $\boldsymbol{A}_s\boldsymbol{1}/\Vert\boldsymbol{A}_s\boldsymbol{1}\Vert$, which is exactly the Rake-DS beamformer. This is analogous to the behavior in the non-raking case (6.12).

6.4.4 Max-UDR Raking

Finally, it is interesting to investigate what happens if we choose the weights that optimize the perceptually motivated UDR [25, 127]. The UDR expresses the fact that adding early reflections (up to 50 ms in the RIR) is as good as adding the energy to the direct sound, as far as speech intelligibility goes. The useful signal is a *coherent* sum of the direct and early reflected speech energy, so that

$$\mathsf{UDR} = \frac{\mathbb{E}\sum_{k=0}^{K}\lvert\boldsymbol{w}^*\boldsymbol{a}(\boldsymbol{s}_k)x\rvert^2}{\mathbb{E}\left\lvert\sum_{k=0}^{K'}\boldsymbol{w}^*\boldsymbol{a}(\boldsymbol{q}_k)z + \boldsymbol{w}^*\boldsymbol{n}\right\rvert^2}. \tag{6.23}$$

In applications K is rarely large enough to cover all the reflections occurring within 50 ms, simply because it is too optimistic to assume we know all the corresponding image sources. Therefore, (6.23) typically underestimates the UDR.

Alas, because (6.23) is specified in the frequency domain, it is challenging to control whether the reflections in the numerator arrive before or after the direct sound, so optimizing this criterion may lead to pre-echoes. Nevertheless, it is interesting to analyze it for the sake of comparison with time-domain raking formulations. It also provides a meaningful metric for evaluation of the raking algorithms presented in this chapter.

To compute the Rake-Max-UDR weights, we solve the following program

$$\underset{\boldsymbol{w}\in\mathbb{C}^M}{\text{maximize}} \frac{\mathbb{E}\sum_{k=0}^{K} |\boldsymbol{w}^* \boldsymbol{a}(\boldsymbol{s}_k)x|^2}{\mathbb{E}\Big|\boldsymbol{w}^* \sum_{k=0}^{K'} \boldsymbol{a}(\boldsymbol{q}_k)z + \boldsymbol{w}^*\boldsymbol{n}\Big|^2}. \tag{6.24}$$

This amounts to maximizing a particular generalized Rayleigh quotient,

$$\frac{\boldsymbol{w}^* \boldsymbol{A}_s \boldsymbol{A}_s^* \boldsymbol{w}}{\boldsymbol{w}^* \boldsymbol{K}_{nq} \boldsymbol{w}}. \tag{6.25}$$

The maximum of this expression is found as the largest eigenvalue in the generalized eigenvalue problem

$$\boldsymbol{A}_s \boldsymbol{A}_s^* \boldsymbol{w} = \lambda \boldsymbol{K}_{nq} \boldsymbol{w}, \tag{6.26}$$

and it is achieved by the corresponding generalized eigenvector ,

$$\boldsymbol{w}_{\text{R-UDR}} = \boldsymbol{w}_{\max}. \tag{6.27}$$

6.4.5 SINR Gain from Raking

Intuitively, if we have multiple sources of the desired signal scattered in space, and we account for it in the design, we should do at least as well as when we ignore the image sources. Let us see how large the gain can be for the Rake-Max-SINR beamformer. We have that

$$\mathsf{SINR} = \sigma_x^2 (\boldsymbol{A}_s \boldsymbol{1})^* \boldsymbol{K}_{nq}^{-1} (\boldsymbol{A}_s \boldsymbol{1}). \tag{6.28}$$

Intuitively, the larger the norm of $\boldsymbol{A}_s\boldsymbol{1}$, the better the SINR as $\boldsymbol{K}_{nq}$ is positive.[4] To explicitly see if there is any gain in using the acoustic rake receiver, we should compare the standard Max-SINR beamformer with the Rake-Max-SINR, *e.g.*, we should evaluate

$$\frac{\big(\sum_k \boldsymbol{a}(\boldsymbol{s}_k)\big)^* \boldsymbol{K}_{nq}^{-1} \big(\sum_k \boldsymbol{a}(\boldsymbol{s}_k)\big)}{\boldsymbol{a}(\boldsymbol{s}_0)^* \boldsymbol{K}_{nq}^{-1} \boldsymbol{a}(\boldsymbol{s}_0)}. \tag{6.29}$$

One possible interpretation of (6.29) is that we ask whether the steering vectors $\boldsymbol{a}(\boldsymbol{s}_k)$ sum coherently or they cancel out.

To answer this, assume that $\boldsymbol{s}_k$, $0 \leqslant k \leqslant K$ are the desired sources (true and image), and let $\beta \overset{\text{def}}{=} \sum_{k=1}^{K} (\alpha_k/\alpha_0)^2$, where α_k is the strength of the source $\boldsymbol{s}_k$ received by the array. Then

$$\mathbb{E}\left(\left\| \sum_{k=0}^{K} \boldsymbol{a}(\boldsymbol{s}_k) \right\|^2\right) \approx (1+\beta)\, \mathbb{E}(\|\boldsymbol{a}(\boldsymbol{s}_0)\|^2), \tag{6.30}$$

that is, we can expect an increase in the output SINR approximately by a factor of $(1+\beta)$ when using the Rake-Max-SINR beamformer. This statement is made precise in Theorem 6.1 in Appendix 6.A. It holds when $\boldsymbol{K}_{nq}$ has eigenvalues of similar magnitude, which is typically not the case in the presence of interferers. However, we show in Section 6.6 that with interferers present, the gains actually increase.

A couple of remarks are in order:

---

[4] It is not difficult to contrive counterexamples, but generically SINR will grow with $\|\boldsymbol{A}_s \boldsymbol{1}\|$.

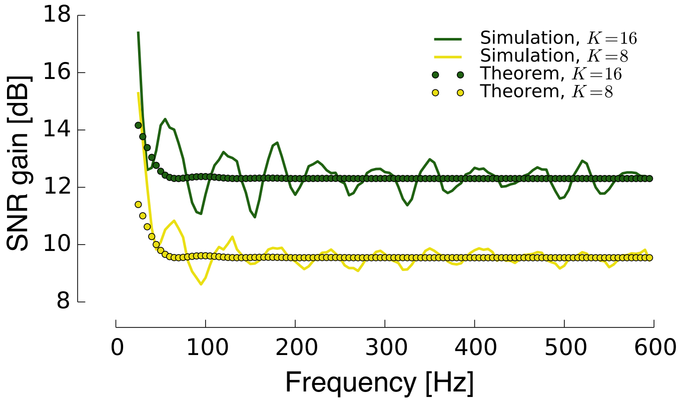


Figure 6.5: Comparison of the simulated SNR gains and the theoretical prediction from Theorem 6.1 for $K = 8$, and $K = 16$. The theoretical prediction of the gain is $10\log_{10}(8+1) \approx 9.54$ for $K = 8$, and $10\log_{10}(16+1) \approx 12.30$ for $K = 16$.

(i) This result is in expectation; it says that on average, the SINR will increase by a factor of $(1+\beta)$. In the worst case, the steering vectors $\boldsymbol{a}(\boldsymbol{s}_k)$ can even cancel out so that the SINR decreases. But the numerical experiments suggest that this is very rare in practice, and we can on the other hand observe large gains.

(ii) We see that summing the phasors in $a_m(\boldsymbol{s}_k)$ behaves as a two-dimensional random walk. It is known that the root-mean-square distance of a 2D random walk from the origin after n steps is $\sqrt{n}$ [140].

(iii) Due to the far-field assumption in Theorem 6.1, the attenuations α_k are assumed to be independent of the microphones; in reality they do depend on the source locations. However, they also depend on a number of additional factors, for example wall attenuations and radiation patterns of the sources. Therefore, for simplicity, we consider them to be independent. One can verify that this assumption does not change the described trend.

It is reassuring to observe the behavior suggested by (6.30) in practice. Figure 6.5 shows the comparison of the prediction by Theorem 6.1 with the SNR gains observed in simulated rooms. In this case, we are comparing the pure SNR gain for white noise, without interferers. To generate Figure 6.5, we randomized the location of the source inside the rectangular room. For simplicity we fixed the signal power as received by the microphones to the same value for all the image sources, so that the expected gain is $K+1$ in the linear scale. The obtained curves agree near-perfectly with the prediction of Theorem 6.1. As we will see, with interferers the gains become larger.

6.5 Comments on Finding and Tracking the Echoes

Thus far we have assumed that the locations of the image sources are known. In this section we briefly describe some methods to localize them when they are *a priori* unknown. We assume that we can localize the true source, or at least one image source. Combined with the knowledge of the room geometry, this suffices to find the locations of other image sources [159].

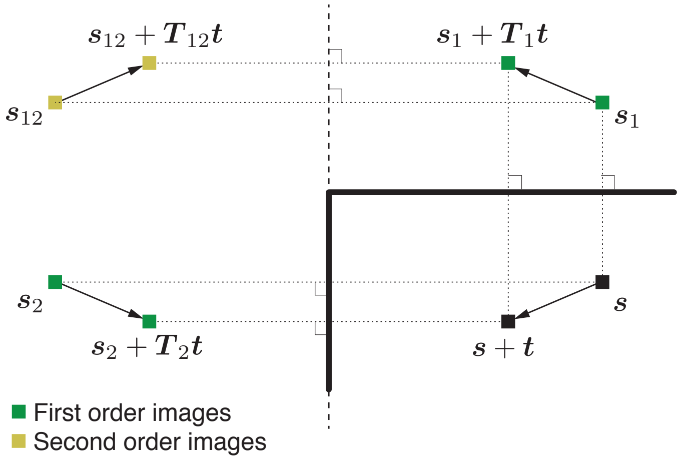


Figure 6.6: Illustration of image source tracking in rectangular geometries.

6.5.1 Known Room Geometry

In many cases, for example for fixed deployments, the room geometry is known. This knowledge could be obtained at the time of the deployment, or from blueprints. In most indoor environments, we encounter a large number of planar reflectors. These reflectors correspond to image sources. Using the tools from Chapter 2 (Section 2.3.4), we can easily compute the image source locations.

6.5.2 Acoustic Image Source Localization

When the room geometry is not known, it is possible to estimate it using the same array that we use for beamforming. We can do it by employing the methods in Chapter 3 in a calibration phase.

But to design an ARR, we do not really need to know how the room looks like; we only need to know where the major echoes are coming from. One possible approach is to locate the image sources in the initial calibration phase, and then track their movement by tracking the true source.

We propose a tracking rule that leverages the knowledge of the displacement of the true source. Again with reference to Figure 6.6, we can state the following simple proposition:

Proposition 6.1

Suppose that the room has only right angles so that the walls are parallel with the coordinate axes. Let the source move from $\boldsymbol{s}$ *to* $\boldsymbol{s} + \boldsymbol{t}$*. Then any image source* $\boldsymbol{s}_k$*, moves to a point given by*

$$\boldsymbol{s}_k + \boldsymbol{T}\boldsymbol{t}, \tag{6.31}$$

where $\boldsymbol{T} = \mathrm{diag}(\pm 1, \mp 1)$ *for odd generations, and* $\boldsymbol{T} = \pm \boldsymbol{I}_2$ *for even generations.*

Proof: The proof follows directly from the figure. The displacement of the image source is the same as the displacement of the true source, passed through a series of reflections. Reflection matrices are diagonal matrices with ±1 on the diagonal, and determinant equal to -1, hence the result.

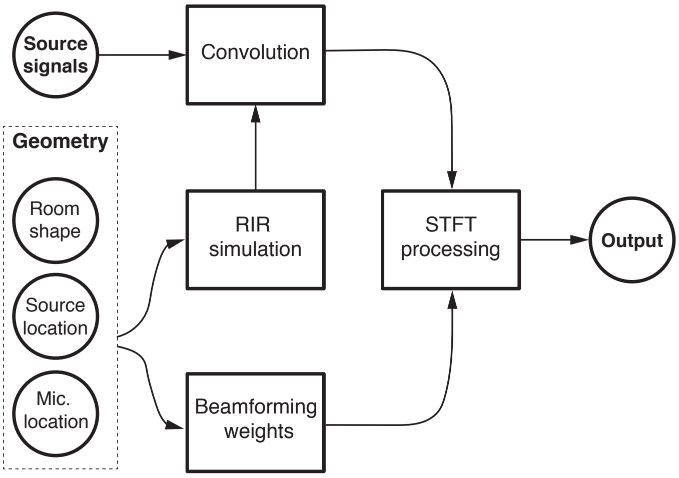


Figure 6.7: Block diagram of the simulation setup used for numerical experiments.

■

The usefulness of this proposition is that it gives us a tool to track the image sources even when we do not know the room geometry (as long as it has right angles). A possible use scenario is to start with a calibration procedure with a controlled source, and perform the echo sorting to find multiple image sources. Then if possible, we assign to each image source a generation (this is in fact a by-product of echo sorting), or we try different hypotheses using Proposition 6.1, and choose the one that maximizes the output SINR.

6.6 Numerical Experiments

In this section, we validate the described theoretical results through numerical experiments. First, we analyze the beampatterns produced by the ARR; second, we evaluate the SINR for various beamformers as a function of the number of image sources used in weight computation; and third, we evaluate the PESQ metric [178]. Finally, we show spectrograms that reveal visually the improved interferer and noise suppression achieved by the ARR.

6.6.1 Simulation Setup

We use a simple room acoustic framework written in Python, that relies on Numpy and Scipy for matrix computations [161]. We limit ourselves to 2D geometry and rectangular rooms. In all experiments, the sampling frequency F_s was set to 8 kHz. An overview of the simulation setup is shown in Figure 6.7.

Starting from the room geometry and the positions of the sources and microphones, we first compute the locations of all images sources up to ten generations. The reflectivity of the walls is fixed to 0.9. The RIR between the source $\boldsymbol{s}_0$ and the microphone $\boldsymbol{r}_m$ is convolved with an ideal

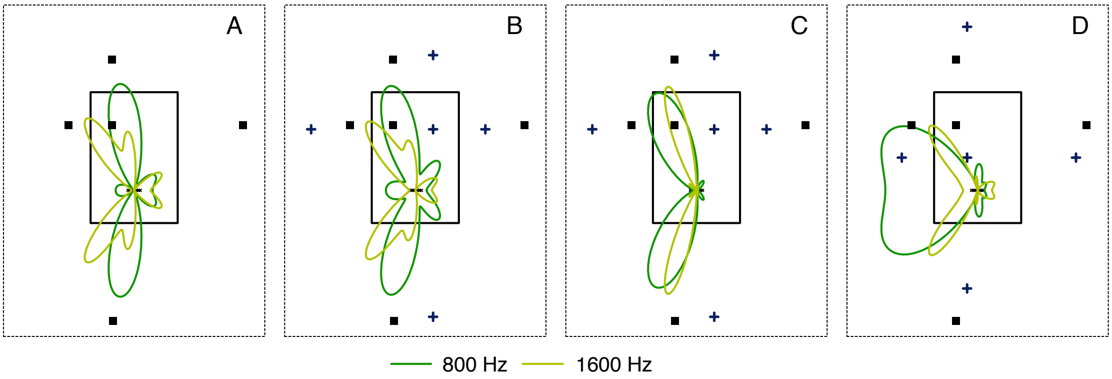


Figure 6.8: Beam patterns in different scenarios. The rectangular room is 4 by 6 meters and contains a source of interest (■) and an interferer (+) ((B), (C), (D) only). The first order image sources are also displayed. The weight computation of the beamformer includes the direct source and the first order image sources of both desired source and interferer (when applicable). (A) Rake-Max-SINR, no interferer, (B) Rake-Max-SINR, one interferer, (C) Rake-Max-UDR, one interferer, (D) Rake-Max-SINR, interferer is in direct path.

low-pass filter in the continuous domain and then sampled at the sampling frequency F_s,

$$\widetilde{a}_m(\boldsymbol{s}_0)[n] = \sum_{k=0}^{K} \frac{\alpha_k}{4\pi\|\boldsymbol{r}_m - \boldsymbol{s}_k\|} \operatorname{sinc}\left(n - F_s\frac{\|\boldsymbol{r}_m - \boldsymbol{s}_k\|}{c}\right), \tag{6.32}$$

where K is the number of image sources considered. We choose the limits of n such that the cardinal sine in (6.32) decays sufficiently to avoid artifacts. The discrete signals from all sound sources are then convolved with their respective RIRs, and added together to obtain the mth microphone's signal.

The beamforming weights are computed in the frequency domain. We use the discrete-time STFT processing with a frame size of $L = 4096$ samples, 50% overlap and zero padding on both sides of the signal by $L/2$. A real fast Fourier transform of size $2L$ and a Hann window are used in the analysis. By exploiting the conjugate symmetry of the real FFT we only need to compute $L+1$ beamforming weights, one for every positive frequency bin. The length L is dictated by the length of the beamforming filters in the time-domain and was set empirically to avoid any aliasing in the filter responses. The output signal was synthesized using the conventional overlap-add method [188].

6.6.2 Results

Beampatterns We first inspect the beampatterns produced by the Rake-Max-SINR and Rake-Max-UDR beamformers for different source-interferer placements. We consider a 4 m × 6 m rectangular room with a source of interest at $(1\text{ m}, 4.5\text{ m})$ and a linear microphone array centered at $(2\text{ m}, 1.5\text{ m})$, parallel to the x-axis. Spacing between the microphones was set to 8 cm. In Figure 6.8, we show the beampatterns for four different configurations of the source and the interferer. We consider a scenario without an interferer, one with an interferer placed favorably at $(2.8\text{ m}, 4.3\text{ m})$, and finally one where the interferer is placed half-way between the desired source and the array at $(1.5\text{ m}, 3.0\text{ m})$.

The last scenario is the least favorable. Interestingly, we can observe that the Rake-Max-SINR beampattern adjusts by completely ignoring the direct path, and steering the beam towards the echoes of the source of interest. This is validating the intuition that we can "hear behind an interferer by listening for the echoes". Note that such a pattern cannot be achieved by a beamformer that only takes into account the direct path. We further note that, while the beampatterns only show the magnitude of the beamformer's response, the phase plays an important role with multiple sources present.

SINR Gains from Raking In the experiments in this subsection, we set the power of the desired source and of the interferer to be equal, $\sigma_x^2 = \sigma_z^2 = 1$. The noise covariance matrix is set to $10^{-3} \cdot \boldsymbol{I}_M$. We use a circular array of $M = 12$ microphones with a diameter of 30 cm, and randomize the position of the desired source and the interferer inside the room. The resulting curves show median performance out of 20000 runs.

Figure 6.9A shows output SINR for different beamformers. The one-forcing beamformer is left out because it performs poorly in terms of SINR, as predicted earlier. Clearly, the Rake-Max-SINR beamformer outperforms all others. The output SINR for beamformers using only the direct path (Max-SINR and DS) remains approximately constant. The UDR is plotted against the number of image sources for various beamformers in Figure 6.9B. The Rake-Max-UDR beamformer performs well in terms of the two measures; however, its output is perceptually unpleasing due to audible pre-echoes; in informal listening tests, the Rake-Max-SINR beamformer did not produce such artifacts. It is interesting to note that the Rake-Max-SINR also performs well in terms of the UDR. Similar SINR gains to those in 6.9A are observed in Figure 6.9C over a range of frequencies. It is therefore justified to extrapolate the results at one frequency in Figure 6.9A to wideband SINR.

Evaluation of Speech Quality We complement the informal listening tests and the evaluation of SINR and UDR with extensive simulations to asses the improvement in speech quality achieved by acoustic raking. We simulate a room with two sources—a desired source and an interferer—and compare the outputs of the Rake-DS, Rake-Max-SINR, and Rake-Max-UDR as a function of the number of image sources used to design the beamformers.

The same number of image sources is used for the target and interferer ($K = K'$). The performance metric used is PESQ [178]. In particular, we use the reference implementation described by the the ITU P.862 Amendment 2 [95]. PESQ compares the reference signal with the degraded signal and predicts the perceptual quality of the latter as it would be measured by the mean opinion score (MOS) value, on a scale from 1 to 4.5.

We consider the same room and microphone array setting as before (see Figure 6.8A). The desired and the interfering sources are placed uniformly at random in a rectangular area with the lower left corner at $(1\text{ m}, 2.5\text{ m})$ and upper right corner at $(3\text{ m}, 5\text{ m})$. To limit the experimental variation, the speech samples attributed to the sources are fixed throughout the simulation. The two sources start reproducing speech at the same time and approximately overlap for the total duration of the speech samples. The signals are normalized to have the same power at the source. We added white Gaussian noise to the microphone signals, with power chosen so that the SNR of the direct sound for the desired source is 20 dB at the center of the microphone array. All signals are high-pass filtered with a cut-off frequency of 300 Hz. The reference for all PESQ results is the direct path of the target signal as measured at the center of the array $(2\text{ m}, 1.5\text{ m})$.

The median PESQ measure of 10000 Monte Carlo runs, given in raw MOS, is shown in Figure 6.9D. The median PESQ of the degraded signal measured at the center of the array

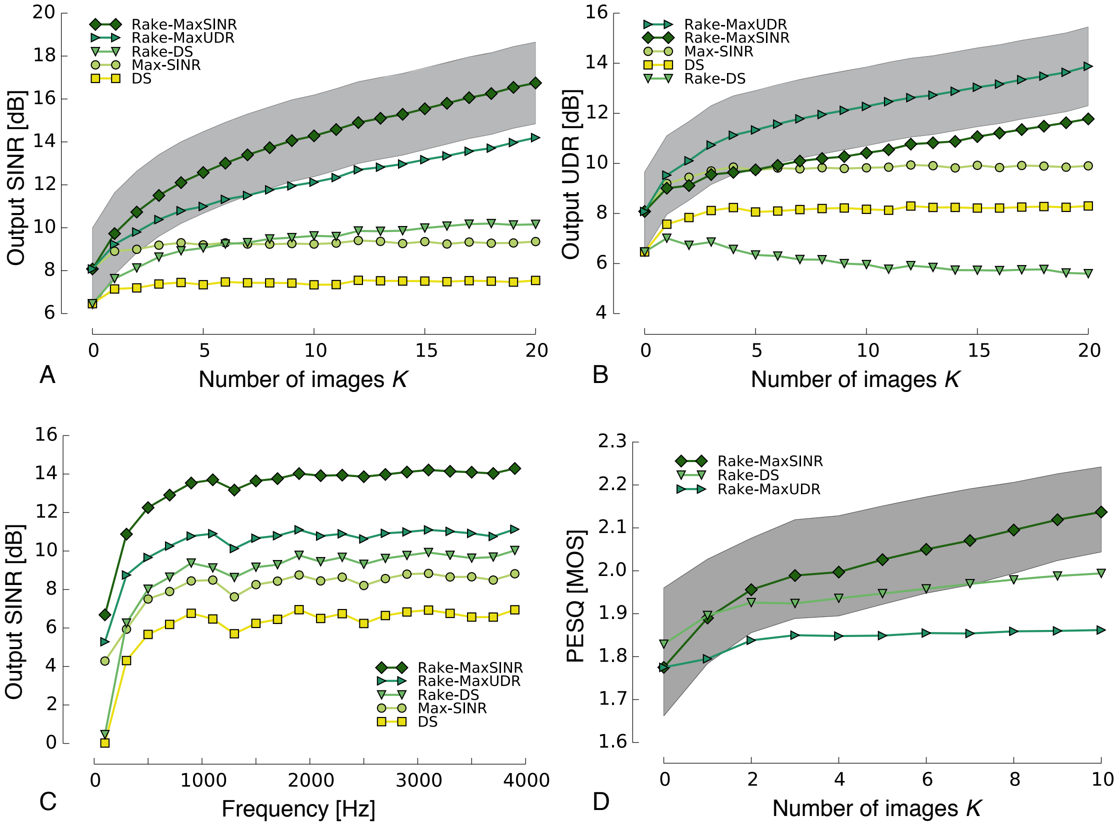


Figure 6.9: Median output SINR (A) and UDR (B) plotted against the number of image sources used in the design for different beamformers, at a frequency $f = 1$ kHz. The shaded area contains the Rake-Max-SINR output SINRs for 50% of the 20000 Monte Carlo runs in (A) and the Rake-Max-UDR output UDR for 50% of the 20000 Monte Carlo runs in (B).

before processing was found to be 1.6 raw MOS. When only the direct sound is used (*i.e.*, $K = 0$), all three beamformers yield the same improvement of about 0.2 raw MOS. We observe that Rake-DS is marginally better than the other beamformers. Using any number of echoes in addition to the direct sound results in larger MOS for all beamformers. When more than one image source is used, the Rake-Max-SINR beamformer always yields the largest MOS, with up to 0.5 MOS gain when using 10 images sources.

It is worth mentioning that in the beamformer design, we do not assume that we know the spectrum of the source or the interferer—we design as if it was flat. Thus the interferer acts as a strong source of colored, spatially correlated, non-stationary noise, spectrally mismatched with the designed beamformer. There is another source of model mismatch: while the RIRs were computed using hundreds of image sources, we use only up to ten to design the beamformers.

Spectrograms and Sounds Samples Finally, we present the spectrograms for a scenario where we want to focus on a singer in the presence of interfering speech. We consider the same room, source, interferer, and microphone array geometry as in Figure 6.8B.

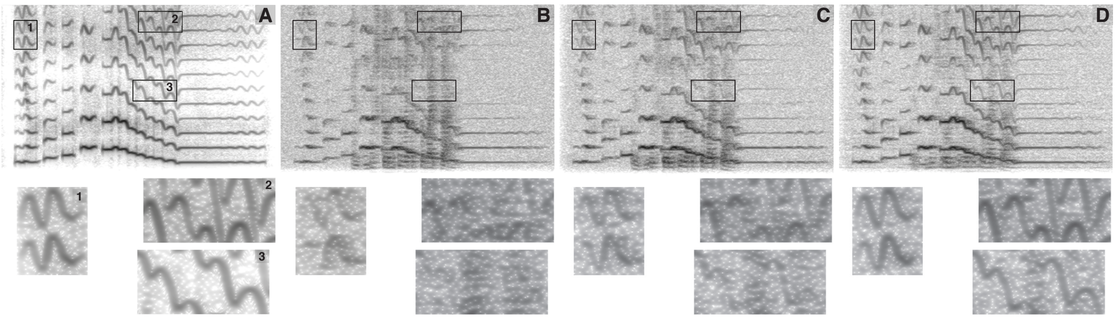


Figure 6.10: Comparison of the conventional Max-SINR and Rake-Max-SINR beamformer on a real speech sample. Spectrograms of (A) clean signal of interest, (B) signal corrupted by an interferer and additive white Gaussian noise at the microphone input, outputs of (C) conventional Max-SINR and (D) Rake-Max-SINR beamformers. Time naturally goes from left to right, and frequency increases from zero at the bottom up to $F_s/2$. To highlight the improvement of Rake-Max-SINR over Max-SINR, we blow-up three parts of the spectrograms in the lower part of the figure. The boxes and the corresponding part of the original spectrogram are numbered in (A). The numbering is the same but omitted in the rest of the figure for clarity.

The source signal is a snippet by a female opera singer (Figure 6.10A), with strongly pronounced harmonics; the interfering signal is a male speech extract. The two signals are normalized to have unit maximum magnitude. We add white Gaussian noise to the microphone signals with power such that the SNR of the direct sound of the desired source is 20 dB at the center of the microphone array. All signals are high-pass filtered with a cut-off frequency of 300 Hz. The Rake-Max-SINR beamformer weights are computed using the direct source and three generations of image sources for both the desired sound source (singing) and the interferer (speech).

The output of the conventional Max-SINR beamformer (Figure 6.10C) is compared to that of the Rake-Max-SINR (Figure 6.10D). We can observe from the spectrogram that the Rake-Max-SINR reduces very effectively the power of the interfering signal at all frequencies, but particularly in the mid to high range. This is true even when the interferer overlaps significantly with the desired signal. Informal listening tests confirm that the Rake-Max-SINR maintains high quality of the desired signal while strongly reducing the interference. The Rake-Max-UDR beamformer provides good interference suppression, but it produces audible pre-echoes that render it unsuitable for speech processing applications. The sound clips can be found online together with the code.

6.7 Time-Domain Formulation

We have seen that the frequency-domain ARR improves over classical beamformers without the need for accurate RIR measurements. We have also seen that the Rake-Max-SINR beamformer improves the mean-opinion score as predicted by the industry standard PESQ.

However, there is potential to do even better if we design the ARR in the time domain. For example, the perceptual quality of the Rake-Max-UDR output is poor because it generates strong pre-echoes. But if we design the ARR in the time domain, we can associate a cost to pre-echoes in the design phase, while leaving the early post-echoes intact.

For the sake of completeness, we explain how to formulate the ARR optimization problems in the time domain. We then propose three raking formulations and discuss preliminary numerical results.

6.7.1 Signal Model

Let us start from the continuous-time domain. The signal at the mth microphone can be written as

$$y_m(t) = \sum_{k=0}^{K} [a_m(\boldsymbol{s}_k, t) * x(t)] + \sum_{k=0}^{K'} [a_m(\boldsymbol{q}_k, t) * z(t)] + e_m(t) \tag{6.33}$$

where $a_m(\boldsymbol{s}_k, t)$ is the channel response between $\boldsymbol{s}_k$ and the $\boldsymbol{r}_m$, and $e_m(t)$ is noise at $\boldsymbol{r}_m$. We assume flat-fading walls, so that the impulse responses are given as the appropriately scaled Green's functions for the wave equation in 3D (*cf.* Chapter 2, Equation (2.56)),

$$a_m(\boldsymbol{s}_k, t) = \frac{\alpha(\boldsymbol{s}_k)}{4\pi\|\boldsymbol{s}_k - \boldsymbol{r}_m\|}\delta\left(t - \frac{\|\boldsymbol{s}_k - \boldsymbol{r}_m\|}{c}\right)$$

where $\boldsymbol{r}_m$ is the position of the mth microphone, c is the speed of sound in air, $\alpha(\boldsymbol{s}_k)$ is the attenuation for the kth image source, and $\delta(t)$ is the Dirac delta distribution. It is convenient to model the channel response as an FIR filter. Thus we low-pass filter it and sample it in the usual way,

$$a_m(\boldsymbol{s}_k, n) = \int_{-\infty}^{\infty} a_m(\boldsymbol{s}_k, u) \operatorname{sinc}(n - F_s\, u)\, \mathrm{d}u = \frac{\alpha(\boldsymbol{s}_k)}{4\pi\|\boldsymbol{s}_k - \boldsymbol{r}_m\|} \operatorname{sinc}\left(n - F_s \frac{\|\boldsymbol{s}_k - \boldsymbol{r}_m\|}{c}\right).$$

Next, we choose a length L_h such that the error incurred by truncating $a_m(\boldsymbol{s}_k, n)$ to L_h samples is negligible,

$$\sum_{|n| \geqslant L_h/2} |a_m(\boldsymbol{s}_k, n)|^2 \approx 0.$$

Using the obtained finite length discretization of the channel responses, we can rewrite (6.33) in matrix form for L_g microphone samples at once:

$$\boldsymbol{y}_m = \sum_{k=0}^{K} \boldsymbol{A}_m(\boldsymbol{s}_k)\boldsymbol{x} + \sum_{k=0}^{K'} \boldsymbol{A}_m(\boldsymbol{q}_k)\boldsymbol{z} + \boldsymbol{e}_m, \tag{6.34}$$

where

$$\begin{aligned}
\boldsymbol{y}_m &= [\ y_m[n] \quad, \ y_m[n-1] \quad, \dots, \quad y_m[n - L_g + 1] \quad]^\top, \\
\boldsymbol{x} &= [\ x[n] \quad, \ x[n-1] \quad, \dots, \quad x[n - L + 1] \quad]^\top, \\
\boldsymbol{z} &= [\ z[n] \quad, \ z[n-1] \quad, \dots, \quad z[n - L + 1] \quad]^\top, \\
\boldsymbol{e}_m &= [\ e_m[n] \quad, \ e_m[n-1] \quad, \dots, \quad e_m[n - L_g + 1] \quad]^\top.
\end{aligned}$$

and $\boldsymbol{A}_m(\boldsymbol{s}_k)$ is the $L_g \times L$ convolution matrix, with L_g being the desired length of the beamforming filter and $L = L_h + L_g - 1$. It is a Toeplitz matrix whose first row is

$$[\ a_m(\boldsymbol{s}_k, 0),\ \dots,\ a_m(\boldsymbol{s}_k, L_h - 1),\ \underbrace{0,\ \dots,\ 0}_{L_g - 1 \text{ times}}\ ], \tag{6.35}$$

and first column is $a_m(\boldsymbol{s}_k, 0)$ padded by $L_g - 1$ zeros.

Stacking all the vectors and matrices indexed by m and dropping the index, we obtain the following compact form of (6.34):

$$\boldsymbol{y} = \boldsymbol{H}_s \boldsymbol{x} + \boldsymbol{H}_q \boldsymbol{z} + \boldsymbol{e},$$

where $\boldsymbol{H}_s = \sum_{k=0}^{K} \boldsymbol{A}(\boldsymbol{s}_k)$ and $\boldsymbol{H}_q = \sum_{k=0}^{K'} \boldsymbol{A}(\boldsymbol{q}_k)$. The mth beamforming filter is $\boldsymbol{g}_m = \left[\; g_m[0], \; \ldots, \; g_m[L_g - 1] \;\right]^\top$ and its output at time n can be written as the inner product $\boldsymbol{g}_m^\top \boldsymbol{y}_m$. Stacking all M filters in a vector, $\boldsymbol{g} = [\; \boldsymbol{g}_0^\top, \; \cdots, \; \boldsymbol{g}_{M-1}^\top \;]^\top$, the sum of all filter outputs can be conveniently computed as $\boldsymbol{g}^\top \boldsymbol{y}$.

We can now concisely write down the responses of the beamformer towards the desired source and towards the interferer,

$$\boldsymbol{u}_s = \boldsymbol{H}_s^\top \boldsymbol{g}, \qquad \boldsymbol{u}_q = \boldsymbol{H}_q^\top \boldsymbol{g}.$$

6.7.2 Time-Domain Rake Beamformers

Minimum Variance Distortionless Response Rake Beamformer A time-domain flavor of the minimum-variance-distortionless-response (MVDR) beamformer [31] is given by[5],

$$\begin{aligned} \underset{\boldsymbol{g}}{\text{minimize}} \quad & \mathbb{E}|\boldsymbol{g}^\top \boldsymbol{y}|^2 \\ \text{subject to} \quad & \boldsymbol{g}^\top \boldsymbol{h}_\tau = 1, \end{aligned} \tag{6.36}$$

where $\boldsymbol{h}_\tau$ is the (τF_s)th column of $\boldsymbol{H}_s$, and τ denotes the delay of the beamformer. Information about the image sources is embedded in $\boldsymbol{h}_\tau$. The constraint promotes unit response towards the desired source, although it does not require the response towards it to be simply a delay. The value of τ should be larger than the time of arrival of the latest arriving echo that we wish to rake.

The optimization problem can be rewritten as

$$\begin{aligned} \underset{\boldsymbol{g}}{\text{minimize}} \quad & \boldsymbol{g}^\top \boldsymbol{R}_{yy} \boldsymbol{g} \\ \text{subject to} \quad & \boldsymbol{g}^\top \boldsymbol{h}_\tau = 1 \end{aligned} \tag{6.37}$$

where $\boldsymbol{R}_{yy}$ is the covariance matrix of $\boldsymbol{y}$,

$$\boldsymbol{R}_{yy} = \boldsymbol{H}_s \boldsymbol{R}_{xx} \boldsymbol{H}_s^\top + \boldsymbol{H}_q \boldsymbol{R}_{zz} \boldsymbol{H}_q^\top + \boldsymbol{R}_{bb},$$

and $\boldsymbol{R}_{xx}$, $\boldsymbol{R}_{zz}$, and $\boldsymbol{R}_{bb}$ are the covariance matrices of $\boldsymbol{x}$, $\boldsymbol{z}$ and the noise. The optimizer is

$$g_{\text{R-MVDR}} = \boldsymbol{R}_{yy}^{-1} \boldsymbol{h}_\tau (\boldsymbol{h}_\tau^\top \boldsymbol{R}_{yy}^{-1} \boldsymbol{h}_\tau)^{-1}.$$

Assuming samples from both sources are independent and identically normally distributed, and that the noise is AWGN, *i.e.*, $\boldsymbol{R}_{xx} = \sigma_x^2 \boldsymbol{I}$, $\boldsymbol{R}_{zz} = \sigma_z^2 \boldsymbol{I}$, and $\boldsymbol{R}_{ee} = \sigma_n^2 \boldsymbol{I}$, (6.37) can be rewritten as

$$\begin{aligned} \underset{\boldsymbol{g}}{\text{minimize}} \quad & \sigma_x^2 \|\boldsymbol{u}_s\|^2 + \sigma_z^2 \|\boldsymbol{u}_q\|^2 + \sigma_n^2 \|\boldsymbol{g}\|^2 \\ \text{subject to} \quad & u_s[\tau] = 1, \\ & \boldsymbol{u}_s = \boldsymbol{H}_s^\top \boldsymbol{g}, \; \boldsymbol{u}_q = \boldsymbol{H}_q^\top \boldsymbol{g}, \end{aligned}$$

---

[5] Although the response is not truly distortionless, we follow the definition of the time-domain MVDR beamformer of Benesty et al [16].

where $u_s[\tau]$ is the τth element of $\boldsymbol{u}_s$. From this form, it is clear that the optimal beamformer will balance distortionless response towards the desired source with interference cancellation and noise suppression. For a fixed L_g, adding more image sources will increase L_h and consequently the number of constraints in the optimization problem, thus taking up degrees of freedom that would otherwise be used for noise suppression.

Finally, using our geometric interpretation it is possible to know precisely how many echoes can be exploited. Because the response is distortionless, the output of the beamformer should be the desired source with a delay τ. This means that we can only exploit the echoes that arrive up to time τ after the direct sound. Using the speed of sound, this translates into a geometrical criterion on which image sources can be included. Concretely, we can use the image sources within a distance of $\|\boldsymbol{s}_0 - \boldsymbol{r}_m\| + c\tau/F_s$ from the microphone array.

Perceptually Motivated Rake Beamformer We already mentioned that early echoes can contribute to the perceived power of the source and improve speedch intelligibility. Lochner and Burger [127] describe precisely how much reverberation is perceptually beneficial. They have found that for speech signals, echoes arriving within 30 ms of the direct sound are fully integrated, while those arriving within 95 ms are still partially integrated. Echoes arriving later than 35 ms are noticeable.

Inspired by these results, we can define the perceptually motivated Rake beamformer subject to the following requirements,

- Minimize the interference and noise power,
- Have zero response before τ (*i.e.* no pre-echoes),
- Have unit response at τ,
- Have zero response after $\tau + \Delta\tau$, where $\Delta\tau \approx 35$ ms.

We can design this beamformer by solving the following quadratic program:

$$\begin{aligned} \underset{\boldsymbol{g}}{\text{minimize}} \quad & \boldsymbol{g}^\top \boldsymbol{K}_{nq} \boldsymbol{g} \\ \text{subject to} \quad & \boldsymbol{g}^\top \widetilde{\boldsymbol{H}}_s = \boldsymbol{\delta}_\tau^\top, \end{aligned} \tag{6.38}$$

where $\boldsymbol{K}_{nq} = \boldsymbol{H}_q \boldsymbol{R}_{zz} \boldsymbol{H}_q^\top + \boldsymbol{R}_{bb}$, the matrix $\widetilde{\boldsymbol{H}}_s$ contains the columns 1 to τ and $\kappa + 1$ to L of $\boldsymbol{H}_s$, and $\boldsymbol{\delta}_\tau$ is the vector with a one at position τ and all other entries zero. Beamforming filters are found as

$$\boldsymbol{g}_{\text{R-P}} = \boldsymbol{K}_{nq}^{-1} \widetilde{\boldsymbol{H}}_s (\widetilde{\boldsymbol{H}}_s^\top \boldsymbol{K}_{nq}^{-1} \widetilde{\boldsymbol{H}}_s)^{-1} \boldsymbol{\delta}_\tau.$$

A similar criterion as for the Rake-MVDR beamformer applies as to which image sources can be used constructively; image sources not farther than $\|\boldsymbol{s}_0 - \boldsymbol{r}_m\| + c(\tau + \kappa)/F_s$ can be exploited.

Maximum SINR Rake Beamformer The signal to interference and noise ratio (SINR) is defined as

$$\mathsf{SINR} = \frac{\mathbb{E}|\boldsymbol{g}^\top \boldsymbol{H}_s \boldsymbol{x}|^2}{\mathbb{E}|\boldsymbol{g}^\top (\boldsymbol{H}_q \boldsymbol{z} + \boldsymbol{b})|^2} = \frac{\boldsymbol{g}^\top \boldsymbol{K}_x \boldsymbol{g}}{\boldsymbol{g}^\top \boldsymbol{K}_{nq} \boldsymbol{g}}, \tag{6.39}$$

where $\boldsymbol{K}_x = \boldsymbol{H}_s \boldsymbol{R}_{xx} \boldsymbol{H}_s^\top$. This quantity can be optimized directly by solving the generalized eigenvalue problem $\boldsymbol{K}_x \boldsymbol{g} = \lambda \boldsymbol{K}_{nq} \boldsymbol{g}$, and the maximizer is given by the generalized eigenvector

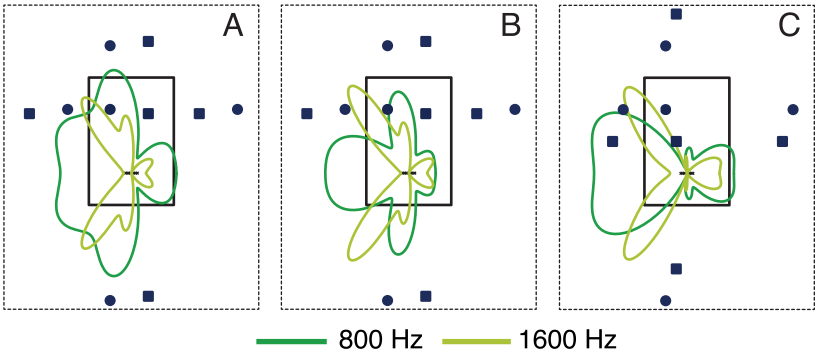


Figure 6.11: Beampatterns of (A) Rake-MVDR, and (B), (C) Rake-Perceptual, in a 4×6 m room containing the desired source (●) and an interferer (■). In (C), the interferer is in the direct path of the desired source. First order image sources are also displayed. The darker/red and light/yellow lines are for 800 Hz and 1600 Hz, respectively.

corresponding to the largest generalized eigenvalue. Unfortunately, this will not yield a practical beamformer—as no constraint is imposed on the response towards the desired source, its signal can be arbitrarily distorted. Nevertheless, this gives an upper bound on achievable SINR.

6.7.3 Numerical Results for TD Beamformers

In this section, we assess the performance of the three rake beamformers described. First, we inspect the beampatterns obtained. Then, the gain of using additional sources is evaluated in terms of output SINR. We use the same simulation setup as for the frequency-domain raking. To simulate room acoustics, we use the image source method with up to 10th order reflections (220 image sources). Samples from both sources are assumed to be zero-mean independent and identically distributed and the noise is AWGN so that

$$\boldsymbol{R}_{xx} = \sigma_x^2 \boldsymbol{I}, \quad \boldsymbol{R}_{zz} = \sigma_z^2 \boldsymbol{I}, \quad \boldsymbol{R}_{bb} = \sigma_n^2 \boldsymbol{I},$$

where $\boldsymbol{I}$ is the identity matrix and $\sigma_x^2 = \sigma_z^2 = 1$.

Beampatterns We consider a 4 by 6 m^2 room with a source of interest at (1 m, 4.5 m) and a linear array of eight microphones equally spaced by 8 cm, parallel to the x-axis and centered at (2 m, 1.5 m) (origin is assumed to be the lower-left corner of the room). The length of the beamforming filters is set to 50 ms which results in $L_g = 400$ at 8 kHz, and the delay τ is set to 20 ms. The noise variance at the microphones is fixed at $\sigma_n^2 = 10^{-7}$.

Beampatterns for both TD-Rake-MVDR and TD-Rake-Perceptual with an interferer placed at (2.8 m, 4.3 m) are shown for 800 Hz and 1600 Hz in Figure 6.11. The diagram in the figure shows the beampatterns for Rake-Perceptual when the interferer is placed in the direct path of the desired source at (1.5 m, 3 m). We observe that similarly as in the frequency-domain case, ARR completely ignores the direct sound and focuses on the reflections.

SINR gain from raking We use the same simulation parameters to evaluate the SINR gain from time-domain raking. Source and interferer positions are chosen independently at random in

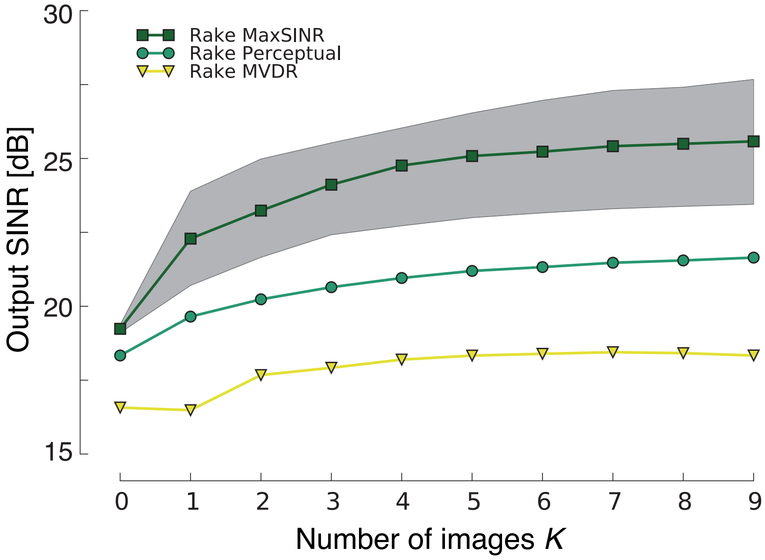


Figure 6.12: Median output SINR computed according to (6.39) against the number of image sources K used in the optimization. The same number of image sources is used for the desired source and the interferer. The ambient noise SNR is fixed to 10 dB with respect to the direct path of the desired source and the center of the microphone array. The grey area contains 50% of the Rake-MaxSINR outcomes.

each Monte Carlo run, and the SINR is computed according to (6.39) for TD-Rake-MVDR, TD-Rake-Perceptual, and TD-Rake-MaxSINR. We mentioned that Rake-MaxSINR is not practical, but it gives an upper bound on the achievable SINR gain. The same number of image sources is used in the design for the source and for the interferer. The noise variance is fixed so that the SNR of the direct path of the desired source is 10 dB at the center of the array. Length of the beamforming filters is set to 30 ms (*i.e.*, $L_g = 240$) and the delay is 20 ms.

The median of 10000 runs is depicted in Figure 6.12. For every beamformer considered, adding more sources results in a net increase in SINR. Adding just the 1st order reflections, or 5 sources, rakes in 1.8 dB improvement for TD-Rake-MVDR and 3 dB SINR improvement for TD-Rake-Perceptual. No beamformer achieves more than 6 dB improvement as shown by the curve for TD-Max-SINR.

6.8 Summary and Conclusion

We investigated the concept of acoustic rake receivers—beamformers that use echoes. Unlike earlier related work, we presented optimal formulations that outperform the delay-and-sum style approaches by a large margin. This is especially true in the presence of interferers, resembling very much the cocktail party scenario—ARRs yield considerable SINR gains.

We have shown that ARRs improve the SINR both theoretically and numerically. Beyond this objective criterion, we have shown for the first time that ARRs improve the subjective quality of speech as predicted by PESQ, proportionally to the number of image sources used. A particularly illustrative example is when the interferer is occluding the desired source—the optimal ARR takes care of this simply by listening to the echoes. Finally, informal listening tests further confirm the improved interference suppression achieved by the ARR.

Perhaps the most important aspect of ongoing work is the design of robust formulations of ARRs. This may involve various heuristics, as well as combinatorial optimization due to the discrete nature of image sources. We expect that the raking beamformers described in this chapter inherit the robustness properties of their classical counterparts. For example, the Rake-DS beamformer is likely to be more robust to array calibration errors than the Rake-Max-SINR beamformer. Furthermore, we expect that taking the image source perspective makes various ARRs more robust to errors in source locations than the schemes that assume the knowledge of the RIR.

Another line of ongoing work investigates the time-domain formulations of the ARRs, with some initial results already available [183]. Time-domain formulations offer better control over whether the echoes appear before or after the direct sound. This provides a more natural framework for optimizing perceptually motivated performance measures, such as the UDR.

6.A Theorem 6.1

We note that the theorem is is stated for a linear array, but the described behavior is universal.

Theorem 6.1

Assume that there are $K+1$ sources located at $\boldsymbol{s}_k = r_k[\cos\theta_k \ \sin\theta_k]^\top$ where $\theta_k \sim \mathcal{U}(0, 2\pi)$ and $r_k \sim \mathcal{U}(a, b)$ are all independent, for some $0 < a < b$ such that the far-field assumption holds. Let $\boldsymbol{A}_s$ collect the corresponding steering vectors for a uniform linear microphone array. Then $\mathbb{E}\|\boldsymbol{A}_s\boldsymbol{1}\|^2 \geqslant (1+\beta)\,\mathbb{E}\|\boldsymbol{a}(\boldsymbol{s}_0)\|^2$, where $\beta = \sum_{k=1}^K (\alpha_k/\alpha_0)^2$, and α_k are attenuations of the steering vectors, assumed independent from the source locations. In fact, $\mathbb{E}(\|\boldsymbol{A}_s\boldsymbol{1}\|^2) = (1+\beta)\,\mathbb{E}(\|\boldsymbol{a}(\boldsymbol{s}_0)\|^2) + O(1/\Omega^3)$.

Proof: Thanks to the far-field assumption, we can decompose the steering vector into a factor due to the array, and a phase factor due to different distances of different image sources. We have that

$$a_m = (\boldsymbol{A}_s\boldsymbol{1})_m = \sum_{k=0}^{K} \alpha_k \mathrm{e}^{-\mathrm{j}\kappa m d \sin\theta_k} \mathrm{e}^{-\mathrm{j}\Omega\delta_k/c}, \tag{6.40}$$

where d is the microphone spacing and $\kappa \overset{\text{def}}{=} \Omega/c$. Without loss of generality we assume that $\delta_k \sim \mathcal{U}(a, b)$. We can further write

$$\begin{aligned}\mathbb{E}|a_m|^2 &= \mathbb{E}\left[\left(\sum_{k=0}^{K} \alpha_k \mathrm{e}^{-\mathrm{j}\kappa m d \sin\theta_k} \mathrm{e}^{-\mathrm{j}\kappa\delta_k}\right)\left(\sum_{\ell=0}^{K} \alpha_\ell \mathrm{e}^{\mathrm{j}\kappa m d \sin\theta_\ell} \mathrm{e}^{\mathrm{j}\kappa\delta_\ell}\right)\right] \\ &= \sum_{k=0}^{K} \alpha_k^2 + \sum_{k\neq\ell=0}^{K} \alpha_k\alpha_\ell \mathbb{E}\left[\mathrm{e}^{\mathrm{j}\kappa m d(\sin\theta_\ell - \sin\theta_k)} \mathrm{e}^{\mathrm{j}\kappa(\delta_\ell-\delta_k)}\right].\end{aligned} \tag{6.41}$$

Invoking the independence for $k \neq \ell$, we compute the above expectation as

$$\mathbb{E}\left[\mathrm{e}^{\mathrm{j}\kappa m d(\sin\theta_\ell - \sin\theta_k)} \mathrm{e}^{\mathrm{j}\kappa(\delta_\ell-\delta_k)}\right] = \frac{2J_0^2(md\kappa)\big[1-\cos(\Delta\kappa)\big]}{(\Delta\kappa)^2}, \tag{6.42}$$

where J_0 denotes the Bessel function of the first kind and zeroth order and $\Delta \overset{\text{def}}{=} b - a$.

Plugging this back into (6.41), we obtain

$$\mathbb{E}|a_m|^2 = \sum_{k=0}^{K} \alpha_k^2 \left(1 + C\frac{2J_0^2(md\kappa)\big[1-\cos(\Delta\kappa)\big]}{(\Delta\kappa)^2}\right), \tag{6.43}$$

where $C = \sum_{k\neq\ell} \alpha_k\alpha_\ell / \sum_k \alpha_k^2$.

Because $|J_0(z)| \leqslant \sqrt{2/(\pi z)} + O(|z|^{-1})$ ([1], Eq. 9.2.1), we see that the expression in brackets is $1 + O(\Omega^{-3})$. Rewriting

$$\sum_{k=0}^{K} \alpha_k^2 = \frac{1}{M}\mathbb{E}\|\boldsymbol{a}(\boldsymbol{s}_0)\|^2 \left(1 + \sum_{k=1}^{K} (\alpha_k/\alpha_0)^2\right) \tag{6.44}$$

concludes the proof. ■

Chapter 7

Norm-Minimizing Generalized Inverses*

7.1 Introduction

Generalized inverses arise in applications ranging from over- and underdetermined linear inverse problems to sparse representations with redundant signal dictionaries. The motivation that led us to writing this chapter comes from a tomographic reconstruction problem we worked on with a start-up company. Because of stringent complexity constraints for a real-time, high frame rate implementation on embedded hardware, precomputed pseudoinverses must be used. In that case, it pays to use "sparse pseudoinverses" instead of the usual Moore-Penrose pseudoinverse. The idea is then to search for an alternative generalized inverse that has a small number of non-zero entries, but still behaves *nicely*. This has since expanded into a general study of norm-minimizing linear generalized inverses.

Linear inverse problems In discrete linear inverse problems, we seek to estimate a signal $\boldsymbol{x}$ from measurements $\boldsymbol{y}$, when they are related by a linear system, $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$, $\boldsymbol{A} \in \mathbb{C}^{m \times n}$. Such problems come in two very different flavors: determined and overdetermined ($m \geqslant n$), and underdetermined ($m < n$). Depending on how we tune the modeling parameters, both cases could occur in the same application. For example, in computed tomography the entries of the system matrix quantify how much the ith ray affects the jth voxel. If we target a coarse reconstruction resolution (less voxels than rays), the system matrix is tall and we deal with an overdetermined system. In this case, we may obtain $\boldsymbol{x}$ from $\boldsymbol{y}$ by applying a generalized (left) inverse, very often the Moore-Penrose pseudoinverse (MPP),[1] When the system is underdetermined ($m < n$), we need to "know what we are looking for" in order to get a meaningful solution; this means that we need a suitable signal model. For most common models (*e.g.* sparsity), the reconstruction of $\boldsymbol{x}$ from $\boldsymbol{y}$ in the underdetermined case is no longer achievable by a linear operator in the style of MPP.

---

[1] The Moore-Penrose pseudoinverse was discovered by Moore in 1920 [153], and later independently by Penrose in 1955 [167].

[*] This chapter is a result of an (as of yet) unpublished joint work with Rémi Gribonval. Parts of the chapter are a joint with Mihailo Kolundžija and Martin Vetterli [52].

Redundant representations In redundant representations, we represent lower-dimensional vectors through higher-dimensional frame and dictionary expansions. The frame expansion coefficients are computed as $\boldsymbol{\alpha} = \boldsymbol{A}^*\boldsymbol{x}$, where $\boldsymbol{A}$ has more columns than rows, and the columns represent the frame vectors. The original signal is then reconstructed as $\boldsymbol{x} = \boldsymbol{D}\boldsymbol{\alpha}$, where $\boldsymbol{D}$ is a *dual frame* of $\boldsymbol{A}$, such that $\boldsymbol{D}\boldsymbol{A}^* = \boldsymbol{I}$. There is a unique correspondence between dual frames and generalized inverses. Different duals lead to different reconstruction properties in terms of resilience to noise and erasures, computational complexity, and other figures of merit [110]. It is therefore interesting to study alternative generalized inverses (alternative to the MPP), in particular those optimal (minimal) according to various criteria; equivalently, it is interesting to study alternative duals.

Generalized inverses beyond the MPP In general, for a matrix $\boldsymbol{A} \in \mathbb{C}^{m\times n}$, there are many different generalized inverses. If $\boldsymbol{A}$ is invertible, they all match. The Moore-Penrose Pseudoinverse (MPP), denoted $\boldsymbol{A}^\dagger$, is special as it optimizes a number of interesting properties. Much of this optimality comes from geometry: for $m < n$, $\boldsymbol{A}^\dagger\boldsymbol{A}$ is an orthogonal projection onto the range of $\boldsymbol{A}^\top$, and this fact turns out to play a key role over and over again. Nevertheless, the MPP is only one of infinitely many generalized inverses, and it is interesting to investigate the properties of others. As the MPP minimizes a particular matrix norm—the Frobenius norm[2]—it seems natural to study alternative generalized inverses that minimize different matrix norms, leading to different optimality properties. Our initial motivation for studying alternative generalized inverses is twofold:

(i) *Efficient computation:* Computing a sparse pseudoinverse gives considerable savings in the speed of computation [52]. Consequently, we may want to compute the sparsest generalized pseudoinverse that is still in some sense stable. A natural object to compute is

$$\mathrm{ginv}_0(\boldsymbol{A}) \overset{\mathrm{def}}{=} \arg\min \|\mathrm{vec}(\boldsymbol{X})\|_0 \ \textit{s.t.}\ \boldsymbol{A}\boldsymbol{X} = \boldsymbol{I}$$

where $\|\cdot\|_0$ counts the total number of nonzero entries of a matrix, which gives the naive complexity of applying $\boldsymbol{X}$ or its adjoint to a vector. Solving the above optimization problem is NP-hard in general (although we will see that for most matrices $\boldsymbol{A}$ the solution is trivial—just invert any full rank minor). Yet, the vast literature establishing equivalence between ℓ^0 and ℓ^1 minimization suggests to replace it by the minimization of the entrywise ℓ^1-norm

$$\mathrm{ginv}_1(\boldsymbol{A}) \overset{\mathrm{def}}{=} \arg\min \|\mathrm{vec}(\boldsymbol{X})\|_1 \ \textit{s.t.}\ \boldsymbol{A}\boldsymbol{X} = \boldsymbol{I}. \tag{7.1}$$

Not only is (7.1) computationally tractable, but it also leads to much better behaved matrices than just inverting a submatrix.

(ii) *Poor man's ℓ^p minimization:* Further motivation comes from an effort to construct a poor man's version of the ℓ^p-minimal solution to an underdetermined set of linear equations $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$. For a general p, the solution to

$$\widehat{\boldsymbol{x}} \overset{\mathrm{def}}{=} \arg\min \|\boldsymbol{x}\|_p \text{ subject to } \boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}, \tag{7.2}$$

cannot be obtained by any linear operator $\boldsymbol{B}$ (see Section 7.2.4). That is, there is no $\boldsymbol{B}$ such that $\boldsymbol{z} \overset{\mathrm{def}}{=} \boldsymbol{B}\boldsymbol{y}$ satisfies $\widehat{\boldsymbol{x}} = \boldsymbol{z}$ for every choice of $\boldsymbol{y}$. The exception is $p = 2$ for which

---

[2] We will see later that it actually minimizes many norms.

the MPP does provide the minimum ℓ^2-norm representation $\boldsymbol{A}^\dagger \boldsymbol{y}$ of $\boldsymbol{y}$; Proposition 7.1 and comments thereafter show that this is indeed the only exception. On the other hand, we can obtain the following bound, valid for any $\boldsymbol{x}$ such that $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{y}$, hence in particular for $\widehat{\boldsymbol{x}}$:

$$\|\boldsymbol{z}\|_p = \|\boldsymbol{B}\boldsymbol{A}\boldsymbol{x}\|_p \leqslant \|\boldsymbol{B}\boldsymbol{A}\|_{\ell^p \to \ell^p} \|\boldsymbol{x}\|_p. \tag{7.3}$$

In particular, if $\boldsymbol{A}\boldsymbol{B} = \boldsymbol{I}$, then $\boldsymbol{z} = \boldsymbol{B}\boldsymbol{y}$ provides an admissible representation $\boldsymbol{A}\boldsymbol{z} = \boldsymbol{y}$, and

$$\|\boldsymbol{z}\|_p \leqslant \|\boldsymbol{B}\boldsymbol{A}\|_{\ell^p \to \ell^p} \|\widehat{\boldsymbol{x}}\|_p. \tag{7.4}$$

This expression suggests that the best generalized inverse $\boldsymbol{B}$ in the sense of minimal worst case ℓ^p-norm blow-up is the one that minimizes $\|\boldsymbol{B}\boldsymbol{A}\|_{\ell^p \to \ell^p}$, motivating the definition of

$$\operatorname{pginv}_p(\boldsymbol{A}) \stackrel{\text{def}}{=} \arg\min \|\boldsymbol{X}\boldsymbol{A}\|_{\ell^p \to \ell^p} \ \ s.t. \ \boldsymbol{A}\boldsymbol{X} = \boldsymbol{I}. \tag{7.5}$$

Objectives The purpose of this chapter is to investigate the properties of generalized inverses defined using various norms. We ask the following questions:

(i) Are there norm families that all lead to the same generalized inverse, thus facilitating the computation?

(ii) Are there specific classes of matrices for which different norms will lead to the same generalized inverse, potentially different from the MPP?

Let us mention that instead of studying *alternative generalized inverses*, we could have decided to take a frame-theoretical perspective [110, 38] and study *alternative dual frames*. These two points of view are equivalent, and our choice is arbitrary.

7.1.1 Prior Art

There are several recent papers in frame theory that study alternative dual frames, or equivalently, generalized inverses. These works concentrate on existence results and explicit constructions of sparse frames and sparse dual frames with prescribed spectra [111, 33]. In particular, Krahmer, Kutyniok, and Lemvig [111] establish sharp bounds on the sparsity of dual frames, showing that generically, for $\boldsymbol{A} \in \mathbb{C}^{m \times n}$ the sparsest dual has $mn - m^2$ zeros.

Unitarily invariant matrix norms are studied in depth by Mirsky [147]. Some important results on the connection between these norms and the MPP are given by Zię̧tak [224]; we comment on these connections in detail in Section 7.4.

Alternative dual frames appear in the literature on sparse representations, for example analysis-based compressed sensing [126]. The idea is that if $\boldsymbol{f}$ is sparse in an overcomplete dictionary $\boldsymbol{D}$, and $\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{f} = \boldsymbol{\Phi}\boldsymbol{D}\boldsymbol{x}$, it is not necessarily the case that $\boldsymbol{D}^*\boldsymbol{f}$ or $\boldsymbol{D}^\dagger\boldsymbol{f}$ is sparse. However, if $\boldsymbol{x}$ is sparse, there must exist *a* dual of $\boldsymbol{D}$ that generates a sparse vector, so the authors propose to optimize not only over $\boldsymbol{f}$, but also over duals of $\boldsymbol{D}$. Another use of generalized inverses different from MPP is when we have some idea about the subspace we want the solution to live in. We can then apply the restricted inverse of Bott and Duffin [20], or its generalizations [146].

MPP can be seen as a minimizer of various cost functions involving the Frobenius norm (as well as of the Frobenius norm itself), over the set of all generalized inverses. The authors in [196] show how to compute *approximate* MPP-like inverses with an additional constraint that the minimizer lives in a particular matrix subspace, and show how to use such matrices to precondition linear systems.

One use of frames is to encode the data before transmitting them over a channel. The channel introduces errors into the data, and the need for robust reconstruction leads to the design of optimal dual frames [122, 129]. Similarly, one can try to compute the best generalized inverse for reconstruction from quantized measurements [118]. Perraudin *et al.* use convex optimization to derive dual Gabor frames with more favorable properties than the canonical dual frame [168], particularly in the sense of time-frequency localization.

In their detailed account of generalized inverses [13], Ben-Israel and Greville use the expression *minimal properties of generalized inverses*, but they primarily concentrate on variations of the square-norm minimality. Additionally, they define a class of non-linear generalized inverses corresponding to various metric projections. We are primarily concerned with generalized inverses that are themselves matrices, but one can imagine various *decoding rules* that search for a vector that satisfies a model, and that is consistent with the measurements [22]. In general, such decoding rules are not linear.

7.1.2 Contributions and the Chapter Outline

We observe that interesting alternative generalized inverses arise through norm minimization. The norm can be placed either on the pseudoinverse $\boldsymbol{X}$ itself, or on the corresponding projection operator $\boldsymbol{X}\boldsymbol{A}$. We study the properties of generalized inverses corresponding to various norms, listed in Table 7.1. The relevant definitions and theoretical results are presented in Section 7.2.

In Section 7.3 we put forward some preliminary results on norm equivalences with respect to norm-minimizing generalized inverses. We summarize the landscape of norm-minimizing generalized inverses tabularly, but also using the *matrix norm cube*—a matrix norm visualization device we developed.

Section 7.4 discusses classes of norms that lead to the MPP. We extend the results of Ziętak on unitarily invariant norms to left unitarily invariant norms. Left-unitary invariance is also relevant when minimizing the norm of the projection operator $\boldsymbol{X}\boldsymbol{A}$, which is in turn relevant for the *poor man's ℓ^p minimization* (Section 7.5.3).

In Section 7.5 we continue the discussion of various norms, by looking at those that almost never yield the MPP. A particular representative of these norms is the entrywise ℓ^1 norm yielding the *sparse pseudoinverse*. We show that minimizing the entrywise ℓ^1-norm of the generalized inverse will always result in a maximally sparse inverse. Here we also talk about poor man's ℓ^p minimization, by discussing generalized inverses that minimize the worst case and the average case ℓ^p blowup. Again, these inverses generally do not coincide with the MPP. However, through numerical simulations we observe that if the forward matrix is a large random matrix, then many of these generalized inverses will be close to each other. In particular, the smallest worst and average case ℓ^p blowup will be achieved by the MPP.

Sections 7.4 and 7.5 talk about norms; in Section 7.6, we concentrate on matrices. We have seen that many norms yield the MPP for all possible input matrices and that some norms generically do not yield the MPP. In Section 7.6 we first discuss a class of matrices for which some of those norms *do* yield the MPP. Particular instances of these classes of matrices are partial Fourier and Hadamard matrices. Next, we exhibit a class of matrices for which many generalized inverses coincide, but *not* with the MPP.

Finally, in Section 7.7 we discuss how to efficiently compute some of the mentioned pseudoinverses. We observe that in some cases the computation simplifies to a vector problem, while in other cases it is indeed a full matrix problem. We propose to use the alternating-direction method of multipliers (ADMM) [164] to compute the generalized inverse, as it can conveniently address both the norms on $\boldsymbol{X}$ and on $\boldsymbol{XA}$.

7.2 Definitions and Known Results (Preliminaries)

Throughout the chapter we assume that all vectors and matrices are over $\mathbb{C}$. We will point out the cases when the result is valid only over the reals. Vectors are all column vectors, and they are denoted with a bold lowercase letters, like $\boldsymbol{x}$. Matrices are denoted by bold capital letters, such as $\boldsymbol{M}$. By $\boldsymbol{M} \in \mathbb{C}^{m \times n}$ we mean that the matrix $\boldsymbol{M}$ has m rows and n columns. The notation $\boldsymbol{I}_m$ stands for the identity matrix in $\mathbb{C}^{m \times m}$; the subscript m will often be omitted. We write $\boldsymbol{m}_j$ for the jth column of $\boldsymbol{M}$, and $\boldsymbol{m}^i$ for its ith row. The conjugate transpose of $\boldsymbol{M}$ is denoted $\boldsymbol{M}^*$. The notation $\boldsymbol{e}_i$ denotes the ith canonical vector.

7.2.1 Generalized Inverses

A generalized inverse of a rectangular matrix is a matrix that has some, but not all properties of the standard inverse of an invertible square matrix. It can be defined for non-square matrices that are not necessarily of full rank.

Definition 7.1 (*Generalized inverse*)

$\boldsymbol{X} \in \mathbb{C}^{n \times m}$ is a generalized inverse of a matrix $\boldsymbol{A} \in \mathbb{C}^{m \times n}$ if it satisfies $\boldsymbol{XA} \boldsymbol{A}$ —

We denote by $\mathcal{G}(\boldsymbol{A})$ the set of all generalized inverses of a matrix $\boldsymbol{A}$.

For the sake of clarity we will primarily concentrate on inverses of underdetermined matrices ($m < n$). As we show in Section 7.3, this choice does not incur a loss of generality. Furthermore, we will often assume that the matrix has full rank: $\operatorname{rank}(\boldsymbol{A}) = m$. In this case, $\boldsymbol{X}$ is the generalized (right) inverse of $\boldsymbol{A}$ if and only if $\boldsymbol{AX} = \boldsymbol{I}_m$.

7.2.2 Correspondence Between Generalized Inverses and Dual Frames

Definition 7.2

A collection of vectors $(\boldsymbol{\phi}_i)_{i=1}^n$ is called a (finite) frame *for $\mathbb{C}^m$ if there exist constants A and B, $0 < A \leqslant B < \infty$, such that*

$$A\|\boldsymbol{x}\|_2^2 \leqslant \sum_{i=1}^{n} |\langle \boldsymbol{x}, \phi_i \rangle|^2 \leqslant B\|\boldsymbol{x}\|_2^2, \tag{7.6}$$

for all $\boldsymbol{x} \in \mathbb{C}^m$.

Definition 7.3

A frame $(\boldsymbol{\psi}_i)_{i=1}^n$ is a dual frame *to $(\boldsymbol{\phi}_i)_{i=1}^n$ if the following holds for any $\boldsymbol{x} \in \mathbb{C}^m$,*

$$\boldsymbol{x} = \sum_{i=1}^{n} \langle \boldsymbol{x}, \boldsymbol{\phi}_i \rangle \boldsymbol{\psi}_i. \tag{7.7}$$

This equation can be rewritten in matrix form as

$$\boldsymbol{x} = \boldsymbol{\Psi}\boldsymbol{\Phi}^{*}\boldsymbol{x}. \tag{7.8}$$

As this must hold for all $\boldsymbol{x}$, we can conclude that

$$\boldsymbol{\Psi}\boldsymbol{\Phi}^{*} = \boldsymbol{I}_{m}, \tag{7.9}$$

and so any dual frame $\boldsymbol{\Psi}$ of $\boldsymbol{\Phi}$ is a generalized left inverse of $\boldsymbol{\Phi}^{*}$. Thus there is a one-to-one correspondence between dual frames and generalized inverses.

7.2.3 Characterization with the Singular Value Decomposition (SVD)

A particularly useful characterization of generalized inverses is through the singular value decomposition (SVD). It has been used extensively to prove theorems in [224, 111] and elsewhere. Consider the SVD of the matrix $\boldsymbol{A}$

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{*}, \tag{7.10}$$

where $\boldsymbol{U} \in \mathbb{C}^{m\times m}$ and $\boldsymbol{V} \in \mathbb{C}^{n\times n}$ are unitary and $\boldsymbol{\Sigma} = \big[\mathrm{diag}(\sigma_1(\boldsymbol{A}), \ldots, \sigma_m(\boldsymbol{A})), \boldsymbol{0}_{m\times(n-m)}\big]$ contains the singular values of $\boldsymbol{A}$ in a non-increasing order. For a matrix $\boldsymbol{X}$, let $\boldsymbol{M} \overset{\mathrm{def}}{=} \boldsymbol{V}^{*}\boldsymbol{X}\boldsymbol{U}$. Then it follows from Definition 7.1 that $\boldsymbol{X}$ is a generalized inverse of $\boldsymbol{A}$ if and only if

$$\boldsymbol{\Sigma}\boldsymbol{M}\boldsymbol{\Sigma} = \boldsymbol{\Sigma}. \tag{7.11}$$

Denoting by r the rank of $\boldsymbol{A}$ and setting

$$\boldsymbol{\Sigma}_{\square} = \mathrm{diag}(\sigma_1(\boldsymbol{A}), \ldots, \sigma_r(\boldsymbol{A})), \tag{7.12}$$

we deduce that $\boldsymbol{M}$ must be of the form

$$\boldsymbol{M} = \begin{bmatrix} \boldsymbol{\Sigma}_{\square}^{-1} & \boldsymbol{R} \\ \boldsymbol{S} & \boldsymbol{T} \end{bmatrix}, \tag{7.13}$$

where $\boldsymbol{R} \in \mathbb{C}^{r\times(m-r)}$, $\boldsymbol{S} \in \mathbb{C}^{(n-r)\times r}$, and $\boldsymbol{T} \in \mathbb{C}^{(n-r)\times(m-r)}$ are arbitrary matrices.

For a full-rank $\boldsymbol{A}$, (7.13) simplifies to

$$\boldsymbol{M} = \begin{bmatrix} \boldsymbol{\Sigma}_{\square}^{-1} \\ \boldsymbol{S} \end{bmatrix} \tag{7.14}$$

In the rest of this chapter we restrict ourselves to full-rank matrices, and use the following characterization of the set of all generalized inverses of a matrix

$$\mathcal{G}(\boldsymbol{A}) = \big\{\boldsymbol{X} \;:\; \boldsymbol{X} = \boldsymbol{V}\boldsymbol{M}\boldsymbol{U}^{*} \text{ where } \boldsymbol{M} \text{ has the form (7.14)}\big\}. \tag{7.15}$$

7.2.4 The Moore-Penrose Pseudoinverse (MPP)

The Moore-Penrose Pseudoinverse (MPP) has a special place among generalized inverses, thanks to its various optimality and symmetry properties.

Definition 7.4 (*MPP*)

The Moore-Penrose Pseudoinverse of the matrix $\boldsymbol{A}$ is the unique matrix $\boldsymbol{A}^\dagger$ satisfying

$$\begin{aligned} \boldsymbol{A}\boldsymbol{A}^\dagger\boldsymbol{A} &= \boldsymbol{A}, & (\boldsymbol{A}\boldsymbol{A}^\dagger)^* &= \boldsymbol{A}\boldsymbol{A}^\dagger, \\ \boldsymbol{A}^\dagger\boldsymbol{A}\boldsymbol{A}^\dagger &= \boldsymbol{A}^\dagger, & (\boldsymbol{A}^\dagger\boldsymbol{A})^* &= \boldsymbol{A}^\dagger\boldsymbol{A}. \end{aligned} \tag{7.16}$$

This definition is universal—it holds regardless of whether $\boldsymbol{A}$ is underdetermined or overdetermined, and regardless of whether it is full rank. Under the conditions primarily considered in this chapter ($m < n$, $\operatorname{rank}(\boldsymbol{A}) = m$), we can express the MPP as $\boldsymbol{A}^\dagger = \boldsymbol{A}^*(\boldsymbol{A}\boldsymbol{A}^*)^{-1}$, which corresponds to the particular choice $\boldsymbol{S} = \mathbf{0}_{(n-m)\times m}$ in (7.14). The canonical dual frame $\boldsymbol{\Phi}$ of a frame $\boldsymbol{\Psi}$ is the adjoint $\boldsymbol{\Phi} = [\boldsymbol{\Phi}^\dagger]^*$ of its MPP.

There are several alternative definitions of MPP. One that is often encountered in the literature is as follows (that this definition makes sense will also be clear from the next section):

Definition 7.5

MPP is the unique generalized inverse of $\boldsymbol{A}$ with minimal Frobenius norm.

As we will see in Section 7.4, the MPP can also be characterized as the generalized inverse minimizing other matrix norms.

The MPP has a number of interesting properties. If $\boldsymbol{A} \in \mathbb{C}^{m\times n}$, with $m > n$, and

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{e}, \tag{7.17}$$

we can compute

$$\widehat{\boldsymbol{x}} = \boldsymbol{A}^\dagger\boldsymbol{y}. \tag{7.18}$$

This vector $\widehat{\boldsymbol{x}}$ is what would in the noiseless case generate $\widehat{\boldsymbol{y}} = \boldsymbol{A}\boldsymbol{A}^\dagger\boldsymbol{y}$—the orthogonal projection of $\boldsymbol{y}$ onto the range of $\boldsymbol{A}$. This is also known as the least-squares solution to an inconsistent overdetermined system of linear equations, in the sense that it minimizes the sum of squared residuals over all equations. For uncorrelated, zero-mean errors of equal variance, this gives the best linear unbiased estimator (BLUE) of $\boldsymbol{x}$.

Note that the optimal solution to (7.17) in the MMSE sense (when $\boldsymbol{x}$ and $\boldsymbol{e}$ are considered random) is not given by the MPP, but rather as the Wiener filter [101],

$$\boldsymbol{B}_{\mathsf{MMSE}} = \boldsymbol{C}_x\boldsymbol{A}^*(\boldsymbol{A}\boldsymbol{C}_x\boldsymbol{A}^* + \boldsymbol{C}_n)^{-1}, \tag{7.19}$$

where $\boldsymbol{C}_x$ and $\boldsymbol{C}_n$ are the signal and noise covariance matrices. The MPP is a special case of this formula for $\boldsymbol{C}_n = \boldsymbol{0}$ and $\boldsymbol{C}_x = \boldsymbol{I}$.

In the underdetermined case, $\boldsymbol{A} \in \mathbb{C}^{m\times n}$, $m < n$, applying the MPP yields the solution with the smallest ℓ^2-norm among all vectors $\boldsymbol{x}$ satisfying $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ (among all *admissible* $\boldsymbol{x}$). That is,

$$\|\boldsymbol{A}^\dagger\boldsymbol{A}\boldsymbol{x}\|_2 \leqslant \|\boldsymbol{z}\|_2, \quad \forall\, \boldsymbol{z} \text{ such that } \boldsymbol{A}\boldsymbol{z} = \boldsymbol{A}\boldsymbol{x}. \tag{7.20}$$

To see this, we use the orthogonality of $\boldsymbol{A}^\dagger\boldsymbol{A}$. Note that any vector $\boldsymbol{x}$ can be decomposed as

$$\boldsymbol{A}^\dagger\boldsymbol{A}\boldsymbol{x} + (\boldsymbol{I} - \boldsymbol{A}^\dagger\boldsymbol{A})\boldsymbol{x}, \tag{7.21}$$

where $\boldsymbol{A}^\dagger \boldsymbol{A}$ is the orthogonal projection on range$(\boldsymbol{A}^*)$, and $(\boldsymbol{I} - \boldsymbol{A}^\dagger \boldsymbol{A})$ the orthogonal projection on range$(\boldsymbol{A}^*)^\perp$. Then we have

$$\begin{aligned} \|\boldsymbol{z}\|_2^2 &= \|\boldsymbol{A}^\dagger \boldsymbol{A} \boldsymbol{z} + (\boldsymbol{I} - \boldsymbol{A}^\dagger \boldsymbol{A})\boldsymbol{z}\|_2^2 \\ &= \|\boldsymbol{A}^\dagger \boldsymbol{A} \boldsymbol{z}\|_2^2 + \|(\boldsymbol{I} - \boldsymbol{A}^\dagger \boldsymbol{A})\boldsymbol{z}\|_2^2 \\ &\geqslant \|\boldsymbol{A}^\dagger \boldsymbol{A} \boldsymbol{z}\|_2^2 = \|\boldsymbol{A}^\dagger \boldsymbol{A} \boldsymbol{x}\|_2^2. \end{aligned} \tag{7.22}$$

A natural question to ask is if there are other linear generalized inverses corresponding to $p \neq 2$. The answer is negative: MPP is the only linear generalized inverse that produces solutions to underdetermined linear systems with a minimal ℓ^p norm, as a consequence of the following result on projections:

Proposition 7.1 (*Theorem 3, [157]*)

Let $M \subset \mathbb{C}^n$ be a linear subspace, and denote by E_M an ℓ^p-norm projection onto M,

$$E_M(\boldsymbol{y}) \overset{\text{def}}{=} \arg\min_{\boldsymbol{x} \in M} \|\boldsymbol{x} - \boldsymbol{y}\|_{\ell^p}. \tag{7.23}$$

Then E_M is linear for all M if and only if $n \leqslant 2$ or $p = 2$.

But all solutions to $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ live in $\boldsymbol{A}^\dagger \boldsymbol{y} + \mathcal{N}(\boldsymbol{A})$, so the ℓ^p-minimal one can be written as

$$\boldsymbol{A}^\dagger \boldsymbol{y} + \arg\min_{\boldsymbol{n} \in \mathcal{N}(\boldsymbol{A})} \|\boldsymbol{A}^\dagger \boldsymbol{y} + \boldsymbol{n}\|_{\ell^p} = \boldsymbol{A}^\dagger \boldsymbol{y} + E_{\mathcal{N}(\boldsymbol{A})}(-\boldsymbol{A}^\dagger \boldsymbol{y}). \tag{7.24}$$

7.2.5 Generalized Inverses Minimizing Matrix Norms

An interesting way of generating different generalized inverses is by norm[3] minimization.

Two central definitions of such generalized inverses will be used in this chapter. The generalized inverse of $\boldsymbol{A} \in \mathbb{C}^{m \times n}$, $m < n$ with minimal ν-norm is defined as ($\|\cdot\|_\nu$ is an arbitrary matrix norm)

$$\operatorname{ginv}_\nu(\boldsymbol{A}) \overset{\text{def}}{=} \arg\min_{\boldsymbol{X}} \|\boldsymbol{X}\|_\nu \quad \text{subject to} \quad \boldsymbol{X} \in \mathcal{G}(\boldsymbol{A}).$$

The generalized inverse minimizing the μ-norm of the product $\boldsymbol{X}\boldsymbol{A}$ is defined as

$$\operatorname{pginv}_\mu(\boldsymbol{A}) \overset{\text{def}}{=} \arg\min_{\boldsymbol{X}} \|\boldsymbol{X}\boldsymbol{A}\|_\mu \quad \text{subject to} \quad \boldsymbol{X} \in \mathcal{G}(\boldsymbol{A}).$$

This definition, which is a particular case of the first one with $\|\cdot\|_\nu = \|\cdot \boldsymbol{A}\|_\mu$, will serve when considering $\boldsymbol{X}$ as a poor man's linear replacement for ℓ^p minimization. Strictly speaking, the above-defined pseudoinverses are sets, since the corresponding programs may have more than one solution. We will point out the cases when special care must be taken.

We will treat several families of matrix norms. A matrix norm is any norm on $\mathbb{C}^{m \times n}$.

Entrywise norms. The simplest matrix norm is the p-entrywise norm. It is defined through an isomorphism between $\mathbb{C}^{m \times n}$ and $\mathbb{C}^{mn}$, *i.e.*, it is simply a vector ℓ^p-norm of the concatenated columns.

---

[3] For brevity, we also loosely call "norm" any quasi-norm such as ℓ^p, $p < 1$, as well as the "pseudo-norm" ℓ^0.

Definition 7.6

The p-entrywise norm of $\boldsymbol{M} \in \mathbb{C}^{m \times n}$, where $0 \leqslant p \leqslant \infty$, is given as

$$\|\boldsymbol{M}\|_p \overset{\text{def}}{=} \|\operatorname{vec}(\boldsymbol{M})\|_p. \tag{7.25}$$

A particular entrywise norm is the Frobenius norm associated to $p = 2$.

Induced norms. An important class of norms are the induced norms (or operator norms). To define these norms, we consider $\boldsymbol{M} \in \mathbb{C}^{m \times n}$ as an operator mapping vectors from $\mathbb{C}^n$ (equipped with an ℓ^p-norm) to $\mathbb{C}^m$ (equipped with an ℓ^q-norm).

Definition 7.7

The $\ell^p \to \ell^q$ induced norm of $\boldsymbol{M} \in \mathbb{C}^{m \times n}$, where $0 < p, q \leqslant \infty$, is

$$\|\boldsymbol{M}\|_{\ell^p \to \ell^q} \overset{\text{def}}{=} \sup_{\boldsymbol{x} \neq 0} \frac{\|\boldsymbol{M}\boldsymbol{x}\|_q}{\|\boldsymbol{x}\|_p}. \tag{7.26}$$

It is straightforward to show that this definition is equivalent to $\|\boldsymbol{M}\|_{\ell^p \to \ell^q} = \sup_{\|\boldsymbol{x}\|_p = 1} \|\boldsymbol{M}\boldsymbol{x}\|_q$. Note that while this is usually defined only for proper norms (i.e., with $1 \leqslant p, q \leqslant \infty$) the definition remains valid when $0 < p < 1$ and/or $0 < q < 1$.

Mixed norms (columnwise and rowwise). An interesting case mentioned in the introduction is the $\ell^1 \to \ell^1$ induced norm of $\boldsymbol{X}\boldsymbol{A}$, as it leads to a sort of optimal poor man's ℓ^1 minimization. The $\ell^1 \to \ell^1$ induced norm is a special case of the family of $\ell^1 \to \ell^q$ induced norms, which can be shown to have a simple expression as columnwise mixed norm

$$\|\boldsymbol{M}\|_{\ell^1 \to \ell^q} = \max_{1 \leqslant j \leqslant n} \|\boldsymbol{m}_j\|_q \overset{\text{def}}{=} \|\boldsymbol{M}\|_{|q,\infty|}. \tag{7.27}$$

More generally, one can consider columnwise mixed norms for any p and q:

Definition 7.8

The columnwise mixed norm $\|\boldsymbol{M}\|_{|p,q|}$ is defined as

$$\|\boldsymbol{M}\|_{|p,q|} \overset{\text{def}}{=} \left(\sum_j \|\boldsymbol{m}_j\|_p^q \right)^{1/q} \tag{7.28}$$

with the usual modification for $q = \infty$.

We deal both with column- and row-wise norms, so we introduce a mnemonic notation to easily tell them apart. Thus $\|\cdot\|_{|p,q|}$ denotes columnwise mixed norms, and $\|\cdot\|_{\underline{\overline{p,q}}}$ denotes rowwise mixed norms, defined as follows:

Definition 7.9

The rowwise mixed norm $\|\boldsymbol{M}\|_{\underline{\overline{p,q}}}$ *is defined as*

$$\|\boldsymbol{M}\|_{\underline{\overline{p,q}}} \stackrel{\text{def}}{=} \left(\sum_i \|\boldsymbol{m}^j\|_p^q\right)^{1/q}, \tag{7.29}$$

with the usual modification for $q = \infty$.

Schatten norms. Another classical norm is the spectral norm, which is the $\ell^2 \to \ell^2$ induced norm. It equals the maximum singular value of $\boldsymbol{M}$, and is a special case of a Schatten norm, just as the Frobenius norm which is the ℓ^2-norm of the vector of singular values of $\boldsymbol{M}$. We can also define a general Schatten norm $\|\sigma(\boldsymbol{M})\|_p$ where $\sigma(\boldsymbol{M})$ is the vector of singular values.

Definition 7.10

The Schatten norm $\|\boldsymbol{M}\|_{S^p}$ *is defined as*

$$\|\boldsymbol{M}\|_{S^p} \stackrel{\text{def}}{=} \|\sigma\boldsymbol{M}\|_p, \tag{7.30}$$

where $\sigma\boldsymbol{M}$ *is the vector of singular values.*

As we will see further on, these are special cases of the larger class of unitarily invariant matrix norms.

7.3 Preliminary Results

In this chapter, we concentrate on generalized inverses of fat matrices—matrices with more columns than rows. We first want to show that there is no loss of generality in making this choice. This is clear for minimizing mixed norms and Schatten norms, as for mixed norms we have that

$$\|\boldsymbol{M}\|_{|p,q|} = \|\boldsymbol{M}^*\|_{\underline{\overline{p,q}}}, \tag{7.31}$$

and for Schatten norm we have

$$\|\boldsymbol{M}\|_{S_p} = \|\boldsymbol{M}^*\|_{S_p}. \tag{7.32}$$

It only remains to be shown for induced norms. We can state the following lemma:

Lemma 7.1

Let $1 \leqslant p, q, p^*, q^* \leqslant \infty$ *with* $\frac{1}{p} + \frac{1}{p^*} = 1$ *and* $\frac{1}{q} + \frac{1}{q^*} = 1$. *Then we have the relation* $\|\boldsymbol{M}\|_{\ell^p \to \ell^q} = \|\boldsymbol{M}^*\|_{\ell^{q^*} \to \ell^{p^*}}$.

In other words, all norms we consider on tall matrices can be converted to norms on their fat transposes, and our results applied.

As a corollary we have

$$\|\boldsymbol{M}\|_{\ell^p \to \ell^\infty} = \|\boldsymbol{M}^*\|_{\ell^1 \to \ell^{p^*}} = \|\boldsymbol{M}^*\|_{|p^*,\infty|} \stackrel{\text{def}}{=} \|\boldsymbol{M}\|_{\underline{\overline{p^*,\infty}}} = \max_{1 \leqslant i \leqslant m} \|\boldsymbol{m}^i\|_{p^*}. \tag{7.33}$$

Next, we show that generalized inverses obtained by minimizing columnwise mixed norms always match minimizing an entrywise norm or an induced norm.

Lemma 7.2

Consider $0 < p \leqslant \infty$ *and a full rank matrix* $\boldsymbol{A}$. *For* $0 < q < \infty$, *we have the set equality*

$$\mathrm{ginv}_{|p,q|}(\boldsymbol{A}) = \mathrm{ginv}_{p}(\boldsymbol{A}).$$

For $q = \infty$ *we have* $\|\cdot\|_{|p,\infty|} = \|\cdot\|_{\ell^1 \to \ell^p}$ *and the set inclusion* $\mathrm{ginv}_p(\boldsymbol{A}) \subset \mathrm{ginv}_{|p,\infty|}(\boldsymbol{A}) = \mathrm{ginv}_{\ell^1 \to \ell^p}(\boldsymbol{A})$.

Proof: For $q < \infty$, minimizing $\|\boldsymbol{X}\|_{|p,q|}$ under the constraint $\boldsymbol{X} \in \mathcal{G}(\boldsymbol{A})$ amounts to minimizing $\sum_j \|\boldsymbol{x}_j\|_p^q$ under the constraints $\boldsymbol{A}\boldsymbol{x}_j = \boldsymbol{e}_j$, where $\boldsymbol{x}_j$ is the jth column of $\boldsymbol{X}$ and $\boldsymbol{e}_j$ the jth canonical vector. Equivalently, one can separately minimize $\|\boldsymbol{x}_j\|_p$ such that $\boldsymbol{A}\boldsymbol{x}_j = \boldsymbol{e}_j$. ■

Hence, when considering columnwise norms, we are primarily interested in minimizing $\|\boldsymbol{X}\boldsymbol{A}\|_{|p,q|}$ rather than $\|\boldsymbol{X}\|_{|p,q|}$.

7.3.1 Summary and Visualization of Matrix Norms

We can see that many of the considered norms coincide for particular choices of parameters. For example, the Schatten 2-norm equals the Frobenius norm as well as the entrywise 2-norm. The induced $\ell^2 \to \ell^2$ norm equals the Schatten ∞-norm, while the induced $\ell^1 \to \ell^p$ norm equals the largest column p-norm, that is to say the $|p,\infty|$ columnwise mixed norm. In an effort to capture these equivalences, we constructed a graphical visualization of the treated matrix norms shown on Figure 7.1, where each considered matrix norm is represented as a point.

For certain matrix norms, we prove (*cf.* Corollary 7.2) that $\mathrm{ginv}_\nu(\boldsymbol{A})$ and $\mathrm{pginv}_\nu(\boldsymbol{A})$ always contain the MPP. We also show that for certain matrices with "flat" MPP (*cf.* Theorem 7.6), $\mathrm{ginv}_\nu(\boldsymbol{A})$ contains the MPP for a large class of mixed norms, also those that normally do not yield the MPP. This is the case in particular for partial Fourier matrices.

The main matrix norms we study in this chapter are listed in Table 7.1.

Norm name	**Symbol**	**Definition**	$\boldsymbol{A}^\dagger \in \mathrm{ginv}_\nu(\boldsymbol{A})$	$\boldsymbol{A}^\dagger \in \mathrm{pginv}_\nu(\boldsymbol{A})$
Schatten p-norms	$\|\boldsymbol{M}\|_{S_p}$	$\|\sigma_i(\boldsymbol{M})\|_p$	✓	✓
Entrywise p-norms	$\|\boldsymbol{M}\|_p$	$\|\mathrm{vec}(\boldsymbol{M})\|_p$	✓ $(p = 2)$	✓ $(p = 2)$
Induced norm (operator norm)	$\|\boldsymbol{M}\|_{\ell^p \to \ell^q}$	$\sup\ \{\|\boldsymbol{M}\boldsymbol{x}\|_q,\ \|\boldsymbol{x}\|_p = 1\}$	✓ $(q = 2)$	✓ $(q = 2)$
			$\times\ (q \neq 2)$	$\times\ (q \neq 2)$
Mixed norm (column-wise)	$\|\boldsymbol{M}\|_{|p,q|}$	$\|\{\|\boldsymbol{m}_j\|_p\}_{j=1}^n\|_q$	cf $\|\boldsymbol{M}\|_p\ (q < \infty)$	✓ $(p = 2)$
			cf $\|\boldsymbol{M}\|_{\ell^1 \to \ell^p}\ (q = \infty)$	$\times\ (p \neq 2)$
Mixed norm (row-wise)	$\|\boldsymbol{M}\|_{\underline{\overline{p,q}}}$	$\|\{\|\boldsymbol{m}^i\|_p\}_{i=1}^m\|_q$	$\times$	$\times$

Table 7.1: Summary of matrix norms considered in this chapter.

7.4 Classes of Norms Yielding the Moore-Penrose Pseudoinverse

A particularly interesting property of the MPP is that it minimizes many of the norms in Table 7.1. This property is related to their unitary invariance, and to the geometric interpretation of the MPP.

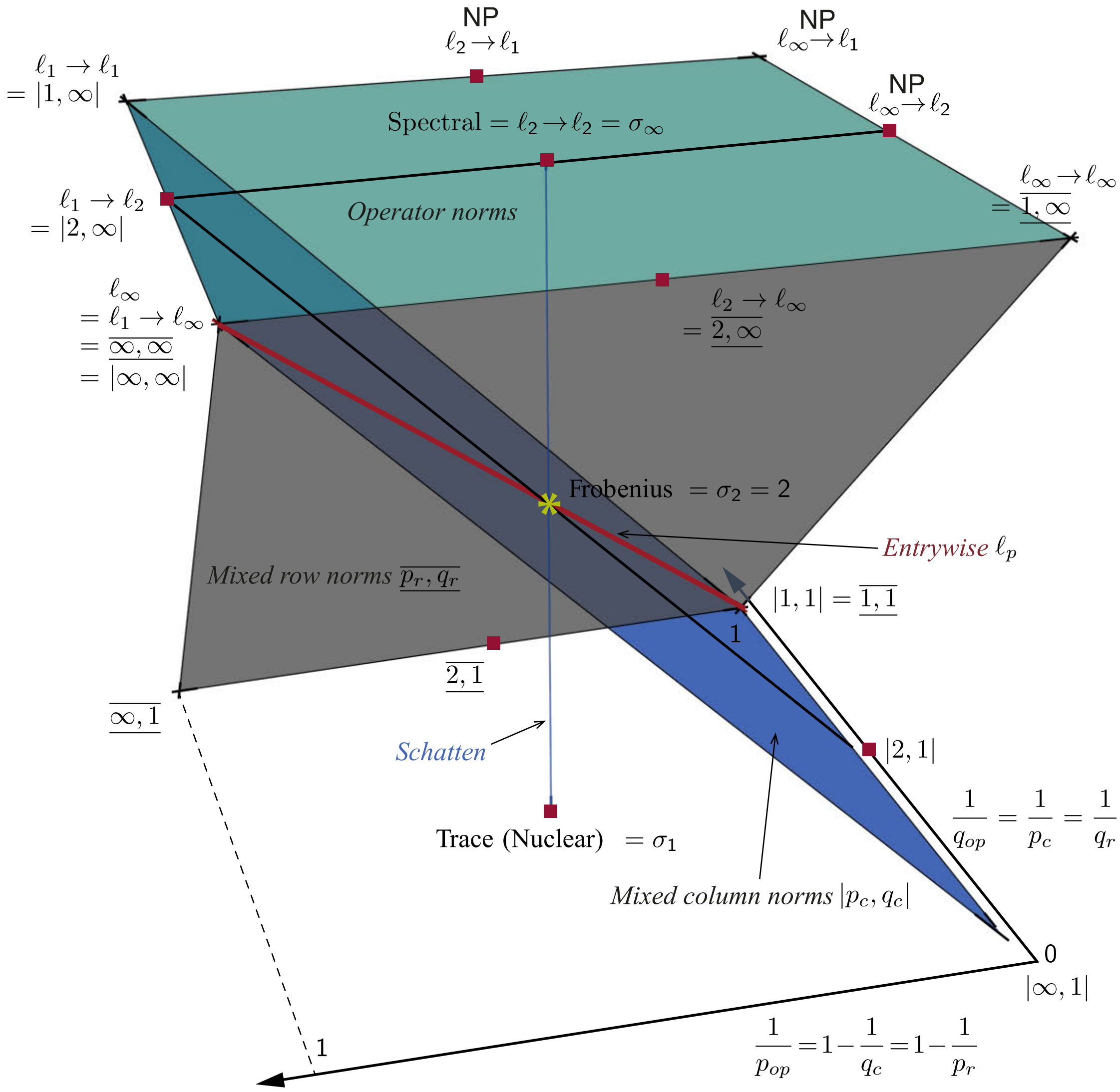

NP
$\ell_2 \to \ell_1$
NP
$\ell_\infty \to \ell_1$
$\ell_1 \to \ell_1$
$= |1, \infty|$
NP
$\ell_\infty \to \ell_2$
Spectral $= \ell_2 \to \ell_2 = \sigma_\infty$
$\ell_\infty \to \ell_\infty$
$= \overline{1, \infty}$
$\ell_1 \to \ell_2$
$= |2, \infty|$
Operator norms
ℓ_∞
$= \ell_1 \to \ell_\infty$
$= \overline{\infty, \infty}$
$= |\infty, \infty|$
$\ell_2 \to \ell_\infty$
$= \overline{2, \infty}$
Frobenius $= \sigma_2 = 2$
Entrywise ℓ_p
Mixed row norms $\overline{p_r, q_r}$
$|1, 1| = \overline{1, 1}$
1
$\overline{2, 1}$
$\overline{\infty, 1}$
Schatten
$|2, 1|$
$\frac{1}{q_{op}} = \frac{1}{p_c} = \frac{1}{q_r}$
Trace (Nuclear) $= \sigma_1$
Mixed column norms $|p_c, q_c|$
0
$|\infty, 1|$
1
$\frac{1}{p_{op}} = 1 - \frac{1}{q_c} = 1 - \frac{1}{p_r}$

Figure 7.1: Graphical representation of matrix norms (the norm cube). The green plane is that of operator norms, the blue one of columnwise mixed norms, and the gray one of rowwise mixed norms. The intersection of rowwise and columnwise mixed norms is shown by the thick red line—these are the entrywise ℓ^p-norms. The vertical blue line is that of Schatten norms, with the nuclear norm S_1 at its bottom, and the spectral norm S_∞ on the top. Among columnwise mixed norms $|p_c, q_c|$, all norms with a fixed value of p_c lead to the same minimizer. Red squares indicate norms ν for which $\mathrm{ginv}_\nu(\boldsymbol{A})$ contains the MPP $\boldsymbol{A}^\dagger$.

7.4.1 Unitarily invariant norms

Definition 7.11 (*Unitarily invariant matrix norm*)

A matrix norm $\|\cdot\|$ is called unitarily invariant if and only if $\|\boldsymbol{UMV}\| = \|\boldsymbol{M}\|$ for any $\boldsymbol{M}$ and any unitary matrices $\boldsymbol{U}$ and $\boldsymbol{V}$.

Unitarily invariant matrix norms are intimately related to symmetric gauge functions [147], defined as vector norms invariant to sign changes and permutations of the vector entries. A theorem by Von Neumann [216, 93] states that any unitarily invariant norm $\|\cdot\|$ is a symmetric gauge function ϕ of the singular values, *i.e.*, $\|\cdot\| = \phi(\sigma(\cdot)) \overset{\text{def}}{=} \|\cdot\|_\phi$. To be a symmetric gauge function, ϕ has to satisfy the following properties [147]:

(i) $\phi(\boldsymbol{x}) \geqslant 0$ for $\boldsymbol{x} \neq 0$,

(ii) $\phi(\alpha\boldsymbol{x}) = |\alpha|\phi(\boldsymbol{x})$,

(iii) $\phi(\boldsymbol{x}+\boldsymbol{y}) \leqslant \phi(\boldsymbol{x}) + \phi(\boldsymbol{y})$,

(iv) $\phi(\boldsymbol{\Pi x}) = \phi(\boldsymbol{x})$,

(v) $\phi(\boldsymbol{\Sigma x}) = \phi(\boldsymbol{x})$,

where $\alpha \in \mathbb{R}$, $\boldsymbol{\Pi}$ is a permutation matrix, and $\boldsymbol{\Sigma}$ is a diagonal matrix with diagonal entries in $\{-1, +1\}$.

Ziętak [224] shows that the MPP minimizes any unitarily invariant norm.

Theorem 7.1 (*Ziętak, 1997*)

Let $\|\cdot\|_\phi$ be a unitarily invariant norm corresponding to a symmetric gauge function ϕ. Then, for any $\boldsymbol{A} \in \mathbb{C}^{m\times n}$, $\|\boldsymbol{A}^\dagger\|_\phi = \min\{\|\boldsymbol{B}\|_\phi : \boldsymbol{B} \in \mathcal{G}(\boldsymbol{A})\}$. If additionally ϕ is strictly monotonic, then the set of minimizers contains a single element $\boldsymbol{A}^\dagger$.

It is interesting to note that in the case of the operator norm, which is associated to the symmetric gauge function $\phi(\cdot) = \|\cdot\|_\infty$, the minimizer is not unique. Ziętak mentions a simple example for rank-deficient matrices, but multiple minimizers are present in the full-rank case too, as illustrated by the following example.

Example 7.1

Let the matrix $\boldsymbol{A}$ *be*

$$\boldsymbol{A} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}. \tag{7.34}$$

Singular values of $\boldsymbol{A}$ *are* $\sigma_1 = \sqrt{3}$ *and* $\sigma_2 = 1$*, and its MPP is*

$$\boldsymbol{A}^\dagger = \boldsymbol{V} \begin{bmatrix} \frac{\sqrt{3}}{3} & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \boldsymbol{U}^* = \frac{1}{3} \begin{bmatrix} 1 & 1 \\ 2 & -1 \\ -1 & 2 \end{bmatrix}. \tag{7.35}$$

Consider now matrices of the form

$$\boldsymbol{A}^\ddagger = \boldsymbol{V} \begin{bmatrix} \frac{\sqrt{3}}{3} & 0 \\ 0 & 1 \\ \alpha & 0 \end{bmatrix} \boldsymbol{U}^*. \tag{7.36}$$

It is readily verified that $\boldsymbol{A}\boldsymbol{A}^\ddagger = \boldsymbol{I}$ *and* $\sigma(\boldsymbol{A}^\ddagger) = \left\{\sqrt{\alpha^2 + \frac{1}{3}},\ 1\right\}$*. Hence, whenever* $0 < |\alpha| \leqslant \sqrt{\frac{2}{3}}$*, we have that* $\|\boldsymbol{A}^\ddagger\|_{S_\infty} = \|\sigma(\boldsymbol{A}^\ddagger)\|_\infty = 1 = \|\boldsymbol{A}^\dagger\|_{S_\infty}$*, and yet* $\boldsymbol{A}^\ddagger \neq \boldsymbol{A}^\dagger$*.*

7.4.2 Left unitarily invariant norms

Another case of particular interest is when the norm is not fully unitarily invariant, but is still unitarily invariant on one side. As we have restricted our attention to fat matrices, we will examine left unitarily invariant norms, because these will conveniently again lead to the MPP. This is the consequence of the following lemma,

Lemma 7.3

Let $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{C}^{n \times m}$ *be defined as*

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{K}_1 \\ \boldsymbol{0} \end{bmatrix}, \quad \boldsymbol{Y} = \begin{bmatrix} \boldsymbol{K}_1 \\ \boldsymbol{K}_2 \end{bmatrix}. \tag{7.37}$$

Then $\|\boldsymbol{X}\| \leqslant \|\boldsymbol{Y}\|$ *for any left unitarily invariant norm* $\|\cdot\|$*.*

Proof: Observe that $\boldsymbol{X} = \boldsymbol{T}\boldsymbol{Y}$ with

$$\boldsymbol{T} = \begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}. \tag{7.38}$$

Now note that

$$\boldsymbol{T} = \frac{1}{2}\begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & -\boldsymbol{I} \end{bmatrix} + \frac{1}{2}\begin{bmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \end{bmatrix}, \tag{7.39}$$

where the two matrices on the right hand side are unitary, and apply the triangle inequality. ■

With this lemma in hand, we can prove the main result for left unitarily invariant norms.

Theorem 7.2

Let $\|\cdot\|$ *be a left unitarily invariant norm, and let* $\boldsymbol{A} \in \mathbb{C}^{m \times n}$ *be full rank with* $m < n$*. Then* $\|\boldsymbol{A}^\dagger\| = \min\ \{\|\boldsymbol{X}\| : \boldsymbol{X} \in \mathcal{G}(\boldsymbol{A})\}$*. If* $\|\cdot\|$ *satisfies a strict inequality in Lemma 7.3 whenever* $\boldsymbol{K}_2 \neq \boldsymbol{0}$*, then the set of minimizers is a singleton* $\{\boldsymbol{A}^\dagger\}$*.*

Proof: Write $\boldsymbol{Y} \in \mathcal{G}(\boldsymbol{A})$ in the form (7.15). By the left unitary invariance and Lemma 7.3

$$\|\boldsymbol{Y}\| = \|\boldsymbol{M}\boldsymbol{U}^*\| = \left\| \begin{bmatrix} \boldsymbol{\Sigma}_{\square}^{-1}\boldsymbol{U}^* \\ \boldsymbol{S}\boldsymbol{U}^* \end{bmatrix} \right\| \geqslant \left\| \begin{bmatrix} \boldsymbol{\Sigma}_{\square}^{-1}\boldsymbol{U}^* \\ \boldsymbol{0} \end{bmatrix} \right\| = \|\boldsymbol{A}^\dagger\|.$$

■

7.4.3 Left unitarily invariant norms on the product operator

We now show that the same phenomenon occurs when minimizing a left unitarily invariant norm of the product $\boldsymbol{X}\boldsymbol{A}$. This simply comes from the observation that if $\|\cdot\|_\mu$ is left unitarily invariant, then so is $\|\cdot\|_\nu = \|\cdot\,\boldsymbol{A}\|_\mu$.

Corollary 7.1

Let $\|\cdot\|$ be a left unitarily invariant norm, and let $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ be full rank with $m < n$. Then $\|\boldsymbol{A}^\dagger\| = \min\ \{\|\boldsymbol{X}\boldsymbol{A}\| : \boldsymbol{X} \in \mathcal{G}(\boldsymbol{A})\}$. If $\|\cdot\|$ satisfies a strict inequality in Lemma 7.3 whenever $\boldsymbol{K}_2 \neq \boldsymbol{0}$, then the set of minimizers is a singleton $\{\boldsymbol{A}^\dagger\}$.

7.4.4 Classical norms leading to the MPP

As a corollary, some large families of norms lead to the MPP.

Corollary 7.2

Let $\boldsymbol{A} \in \mathbb{C}^{m\times n}$ be full rank with $m < n$.

- **Schatten norms:** *for $1 \leqslant p \leqslant \infty$*

  $$\boldsymbol{A}^\dagger \in \mathrm{ginv}_{S_p}(\boldsymbol{A}) \qquad \boldsymbol{A}^\dagger \in \mathrm{pginv}_{S_p}(\boldsymbol{A}).$$

  The considered sets are singletons for $p < \infty$.
  The set $\mathrm{ginv}_{S_\infty}(\boldsymbol{A})$ is not necessarily a singleton; $\mathrm{pginv}_{S_\infty}(\boldsymbol{A})$ is a singleton.

- **Columnwise mixed norms:** *for $1 \leqslant q \leqslant \infty$*

  $$\boldsymbol{A}^\dagger \in \mathrm{ginv}_{|2,q|}(\boldsymbol{A}) \qquad \boldsymbol{A}^\dagger \in \mathrm{pginv}_{|2,q|}(\boldsymbol{A})$$

  The considered sets are singletons for $q < \infty$, but not always for $q = \infty$.

- **Induced norms:** *for $1 \leqslant p \leqslant \infty$*

  $$\boldsymbol{A}^\dagger \in \mathrm{ginv}_{\ell^p\to\ell^2}(\boldsymbol{A}) \qquad \boldsymbol{A}^\dagger \in \mathrm{pginv}_{\ell^p\to\ell^2}(\boldsymbol{A})$$

  The set $\mathrm{ginv}_{\ell^p\to\ell^2}(\boldsymbol{A})$ is not always a singleton for $p \leqslant 2$.

The proof is given in Appendix 7.A.

Remark 1

Let us highlight an interesting consequence of Corollary 7.2: The computation of the $\|\cdot\|_{\ell^\infty\to\ell^2}$ norm is known to be NP-complete [123]. Despite this fact, Corollary 7.2 implies that we can find a solution of an optimization program involving this norm.

7.5 Norms that Almost Never Yield the MPP

After discussing matrix norms whose minimization *always* leads to the MPP, it is interesting to look at norms that generically *never* lead to the MPP. Among the norms that we study, these are columnwise mixed norms with $p \neq 2$, rowwise norms, and induced norms for $q \neq 2$. Let us show this in particular for the columnwise mixed norms $\|\cdot\|_{|p,q|}$ $p \neq 2$. We can state the following result:

Proposition 7.2

> *Let the entries of* $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ *be drawn from some absolutely continuous probability distribution. Then for* $p \neq 2, p \geqslant 1$ *and* $1 \leqslant q \leqslant \infty$ *we have that* $\boldsymbol{A}^\dagger \notin \mathrm{ginv}_{|p,q|}(\boldsymbol{A})$ *with probability one.*

Proof: First note that because the optimization for $\mathrm{ginv}_{|p,q|}(\boldsymbol{A})$ decouples over columns, it is sufficient to show that the ith column of the MPP does not minimize the corresponding column problem for any i. That is,

$$\exists \boldsymbol{x} \in \mathbb{R}^n \text{ such that } \boldsymbol{A}\boldsymbol{x} = \boldsymbol{e}_i \text{ and } \|\boldsymbol{A}^\dagger \boldsymbol{e}_i\|_p^p > \|\boldsymbol{x}\|_p^p \Leftrightarrow \text{MPP is not a minimizer.} \tag{7.40}$$

By integrating the constraint $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{e}_i$ into the cost function, the cost function for the ith column can be written as $f(\boldsymbol{z}) = \|\boldsymbol{A}^\dagger \boldsymbol{e}_i + \boldsymbol{N}\boldsymbol{z}\|_p^p$ (minimization is over $\boldsymbol{z} \in \mathbb{R}^{n-m}$), where columns of $\boldsymbol{N}$ form the orthogonal basis for $\mathcal{N}(\boldsymbol{A})$. The gradient of f is computed as

$$\nabla f(\boldsymbol{z}) = p\boldsymbol{N}^\top \left|\boldsymbol{A}^\dagger \boldsymbol{e}_i + \boldsymbol{N}\boldsymbol{z}\right|^{p-1} \circ \mathrm{sign}(\boldsymbol{A}^\dagger \boldsymbol{e}_i + \boldsymbol{N}\boldsymbol{z}). \tag{7.41}$$

Our goal is now to show that $\nabla f(\boldsymbol{0}) \neq 0$ almost surely. Denoting $\boldsymbol{X} = \boldsymbol{A}^\dagger$, $\boldsymbol{x}_i = \boldsymbol{X}\boldsymbol{e}_i$, and $g(\boldsymbol{x}) = |\boldsymbol{x}|^{p-2} \circ \boldsymbol{x}$, this is equivalent to showing that

$$\exists i \text{ such that } \boldsymbol{N}^\top g(\boldsymbol{x}_i) \neq \boldsymbol{0}, \tag{7.42}$$

or in other words, that $g(\boldsymbol{x}_i) \neq \mathrm{range}(\boldsymbol{A}^\top)$ for at least one i. Note next that $\mathrm{range}(\boldsymbol{A}^\top) = \mathrm{range}(\boldsymbol{X})$, and that while elements of $\boldsymbol{X}$ are not independent, their joint distribution is still absolutely continuous. We can now rephrase our task: For a matrix $\boldsymbol{X}$ whose elements have a continuous distribution, show that $\boldsymbol{g}(\boldsymbol{x}_1) \notin \mathrm{range}(\boldsymbol{X})$. Equivalently, we ask that $(\boldsymbol{I} - \boldsymbol{X}\boldsymbol{X}^\dagger)g(\boldsymbol{x}_1) \neq \boldsymbol{0}$. But entries of $\boldsymbol{X}^\dagger$ are rational functions of the entries of $\boldsymbol{X}$, so the entries of the vector $(\boldsymbol{I} - \boldsymbol{X}\boldsymbol{X}^\dagger)g(\boldsymbol{x}_1) \neq \boldsymbol{0}$ are rational functions whose numerator is a non-zero polynomial in x_{ij} assuming for that p is even. To see that the above polynomial is not identically zero it suffices to find one $\boldsymbol{X}$ for which it is non-zero. Consider the following construction:

$$\boldsymbol{X} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & & \ddots & \\ 0 & 0 & \cdots & 1 \\ 0 & 0 & \cdots & 0 \\ \vdots & & & \\ 2 & 0 & \cdots & 0 \end{bmatrix}, \quad \text{so that} \quad \boldsymbol{X}^\dagger = \begin{bmatrix} \frac{1}{5} & 0 & 0 & 0 & 0 & \cdots & \frac{2}{5} \\ 0 & 1 & \cdots & 0 & 0 & \cdots & 0 \\ & & & & & & \\ 0 & 0 & \ddots & 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 1 & 0 & \cdots & 0 \end{bmatrix}. \tag{7.43}$$

Then it is straightforward to verify that $(\boldsymbol{I} - \boldsymbol{X}\boldsymbol{X}^\dagger)g(\boldsymbol{x}_1) \neq \boldsymbol{0}$ for all $p \neq 2$.

As the joint distribution of x_{ij} is absolutely continuous, and the sets of zeros of these polynomial have measure zero, the probability (with respect to realizations of $\boldsymbol{A}$) that all entries of $(\boldsymbol{I} - \boldsymbol{X}\boldsymbol{X}^\dagger)g(\boldsymbol{x}_1)$ vanish is zero. For a rational $p = q/r$ we can use similar arguments using a reparameterization in terms of $z_{ij} = \mathrm{sign}(x_{ij})|x_{ij}|^{1/r}$. ■

Remark 2

Note that this straightforwardly extends into a proof that the ℓ^p-minimal solution to an underdetermined system of linear equations generically changes with p (in particular it is different from the least-squares solution).

7.5.1 Sparse Pseudoinverse

We now concentrate on ginv, rather than pginv, and in particular on the generalized inverse minimizing the entrywise ℓ^1-norm. We know from Proposition 7.2 that this generalized inverse is in general not the MPP.

The motivation to look specifically at ℓ^1-norm is that one expects it to promote the sparsity of the entries, resulting in what could be called the *sparse pseudoinverse.* The sparse pseudoinverse was studied in [52], where it was shown empirically that the minimizer is indeed a sparse matrix, and that it gives a fast way to solve certain inverse problems. It is convenient to define the specific notation $\operatorname{spinv}(\boldsymbol{A}) = \operatorname{ginv}_1(\boldsymbol{A})$.

The usual first step would be to try and compute $\operatorname{ginv}_0(\boldsymbol{A})$; that is, the sparsest generalized inverse of $\boldsymbol{A}$. This computation is in general not tractable, so we replace $\|\cdot\|_0$ by its convex envelope $\|\cdot\|_1$ [62]. On the other hand, we will show that finding *a* minimizer of $\|\cdot\|_0$ is generically trivial, but usually not a good idea (unlike in the compressed sensing scenario).

The intuitive reasoning is as follows: The constraint $\boldsymbol{A}\boldsymbol{X} = \boldsymbol{I}_m$ is a shorthand notation for m^2 linear equations, so it constraints m^2 degrees of freedom. The matrix $\boldsymbol{X}$ has nm entries, thus we are left with $nm - m^2$ degrees of freedom, which we hope will be set to zero by ℓ^1 minimization. Indeed, we can state the following result:

Theorem 7.3

Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $m < n$. Then there exists $\boldsymbol{X} \in \operatorname{spinv}(\boldsymbol{A})$ that contains at least $mn - m^2$ zero elements.

Remark 3

As a consequence of this theorem, in general the sparse pseudoinverse is not *the MPP.*

In proving this theorem we use a known fact about ℓ^p minimization for $p \leqslant 1$: the set of minimizers always contains an m-sparse point.

Lemma 7.4

Let $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$, where $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ with $m < n$. Then the set of minimizers of $\|\boldsymbol{z}\|_p$, $0 \leqslant p \leqslant 1$, subject to the constraint that $\boldsymbol{A}\boldsymbol{z} = \boldsymbol{y}$, always contains an m-sparse point:

Proof (Theorem 7.3) : Minimization for $\operatorname{spinv}(\boldsymbol{A})$ can be decoupled into ℓ^1 minimizations for every column,

$$\operatorname{spinv}(\boldsymbol{A})_i = \arg\min_{\boldsymbol{A}\boldsymbol{x}=\boldsymbol{e}_i} \|\boldsymbol{x}\|_{\ell^1}. \tag{7.44}$$

By Lemma 7.4, the set of minimizers for every column of $\operatorname{spinv}(\boldsymbol{A})$—solution to (7.44) contains a point with $n - m$ zeros. Because $\operatorname{spinv}(\boldsymbol{A})$ has m columns, there exists a minimizer for $\operatorname{spinv}(\boldsymbol{A})$ with at least $m(n - m)$ zero entries. ■

A corollary of Theorem 7.3 is that, generically, $\operatorname{spinv}(\boldsymbol{A})$ is the sparsest pseudoinverse of $\boldsymbol{A}$. To see this, we invoke a result by Krahmer, Kutyniok and Lemvig [111]:

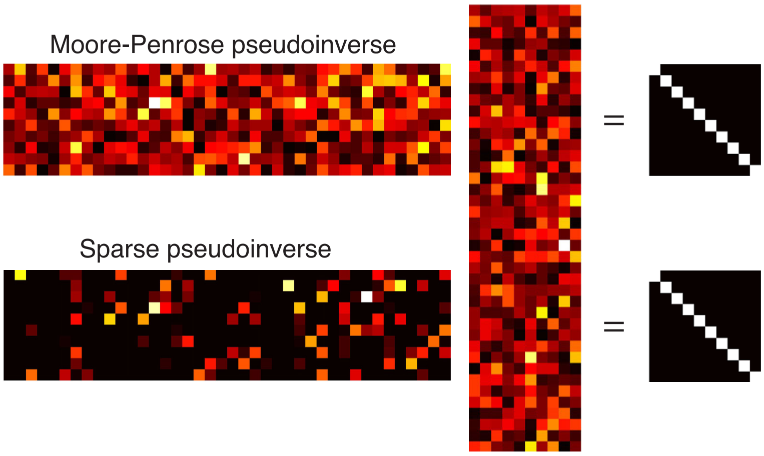


Figure 7.2: Illustration of a sparse pseudoinverse of $\boldsymbol{A} \in \mathbb{R}^{10\times 40}$. Both $\boldsymbol{X}_1$ (upper) and $\boldsymbol{X}_2$ (lower) are right inverses of $\boldsymbol{A}$, but the sparse pseudoinverse $\boldsymbol{X}_2$ has a large number (300 or 75%) of zero entries.

Lemma 7.5 (*[111, Theorem 3.6]*)

Let $\mathcal{F}(m,n)$ be the set of full rank matrices in $\mathbb{C}^{m\times n}$, and denote by $\mathcal{N}(m,n)$ the subset of matrices in $\mathcal{F}(m,n)$ whose sparsest generalized inverse has m^2 non-zeros. Then

(i) Any matrix in $\mathcal{F}(m,n)$ is arbitrarily close to a matrix in $\mathcal{N}(m,n)$,

(ii) The set $\mathcal{F}(m,n) \setminus \mathcal{N}(m,n)$ has measure zero.

In particular, many random matrices (*e.g.* iid Gaussian) will have a sparsest generalized inverse with sparsity m^2. We can then state the following important result on sparse pseudoinverses:

Theorem 7.4

The set of matrices in $\mathcal{F}(m,n)$ for which spinv does not contain the sparsest generalized inverse has measure zero.

Numerical Stability For matrices in a general position (that is, for most matrices), there is a simpler way to obtain a generalized inverse with the minimal number m^2 of non-zeroes—just invert any full-rank $m \times m$ submatrix of $\boldsymbol{A}$. But there is a good reason not to do so in practice: Computing the sparse pseudoinverse is a better idea because of *stability*. Consider for a change an overdetermined inverse problem $\boldsymbol{y} = \boldsymbol{B}\boldsymbol{x} + \boldsymbol{z}$, with $\boldsymbol{B} \in \mathbb{C}^{n\times m}$ where $m < n$, and $\boldsymbol{z}$ is random noise. For any matrix $\boldsymbol{W} \in \mathcal{G}(\boldsymbol{B})$ we have that

$$\mathbb{E}[\|\boldsymbol{W}\boldsymbol{y} - \boldsymbol{x}\|_2^2] = \|\boldsymbol{W}\|_F^2 \sum_{i=1}^{n} \mathbb{E}[|z_i|^2]. \tag{7.45}$$

In noisy situations, we want to control the influence of the noise in the output. We see from the above that this error is dictated by the Frobenius norm of $\boldsymbol{W}$, so it is desirable to use generalized inverses with small Frobenius norms. Alas, for a simple inversion of an $m \times m$ submatrix, this Frobenius norm can be quite large (it is known that the smallest singular value of an $m \times m$ iid random matrix is $\sim m^{-1/2}$ with high probability [180, 69]).

On the contrary, the sparse pseudoinverse as well as other norm-minimizing generalized inverses have controlled Frobenius norms. This is shown for Gaussian random matrices in Figure 7.3 and for Bernoulli random matrices in Figure 7.4. The two figures further show that interesting norm-minimizing generalized inverses behave *nicely* with respect to various matrix norms. This suggests that these generalized inverses will behave in a stable manner with respect to noise in overdetermined inverse problems.

Motivation for Sparse Pseudoinverse (Numerical Experiment) We initially proposed the sparse pseudoinverse to speed up overdetermined tomographic inversion which had to be performed at a fast rate on computation-constrained embedded hardware. Multiplication by the MPP of the system matrix was far too expensive for the purpose. To illustrate the benefits offered by the sparse pseudoinverse, we test the performance of several methods for solving overdetermined systems, including the sparse pseudoinverse.

The upper part of Figure 7.5 shows the performance of various reconstruction methods with a dense forward matrix. We show results for spinv($\boldsymbol{A}$) and sspinv($\boldsymbol{A}$) which is just spinv with hard-thresholded entries. The threshold τ is computed from the empirical CDF of the absolute values of matrix entries shown in the same figure. The number of non-zeros in sspinv is about two times lower than for spinv($\boldsymbol{A}$), and four times lower than for MPP. As we will see spinv and sspinv are particularly interesting in the sparse matrix case. We also include well-known non-linear iterative methods: Kaczmarz [100] and randomized Kaczmarz algorithms [195]. The number of iterations in these algorithms is set to have the same operation count as for the application of sspinv($\boldsymbol{A}$). We sampled 50 realizations of $\boldsymbol{A}$, with entries $a_{ij} \sim \mathcal{U}([0,1])$. For each matrix, a different random input vector was generated, and 50 different iid Gaussian noise vectors were added to the simulated measurements $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$. We can see that at all but the lowest SNRs, the MPP performs the best. But the cost of applying and storing spinv is 2 times lower, and the cost of applying the sspinv 4 times lower. The spinv and sspinv perform similarly up to 30 dB, and both variants of Kaczmarz algorithm perform worse at all but the lowest tested SNRs. This example indicates the merit of computing a sparse pseudoinverse.

In applications like tomography we often encounter sparse forward matrices. This makes the Kaczmarz algorithm attractive, especially for large systems. Surprisingly, we demonstrate that sparse pseudoinvese may be an attractive option even for sparse matrices. We run the simulation like in the previous case, but with the fraction of non-zeros in the forward matrix set to 0.03. This is the sparsity of the model matrix in our practical setup (but with smaller matrix dimensions). The matrix size was 202×101. In the lower half of Figure 7.5, we see that the MPP consistently performs better than spinv at all input SNRs by about 6 dB. Again, it requires at least twice the number of operations. The sparser sspinv performs the same as spinv up to higher SNR values (around 40 dB). But here, sspinv requires 38 times less computation than the MPP. The randomized Kaczmarz algorithm performs the best on the average at low SNRs (below 15 dB), but its performance at higher SNRs is notably inferior to all three linear reconstruction methods. Conventional Kaczmarz has the worst average performance at all tested SNRs.

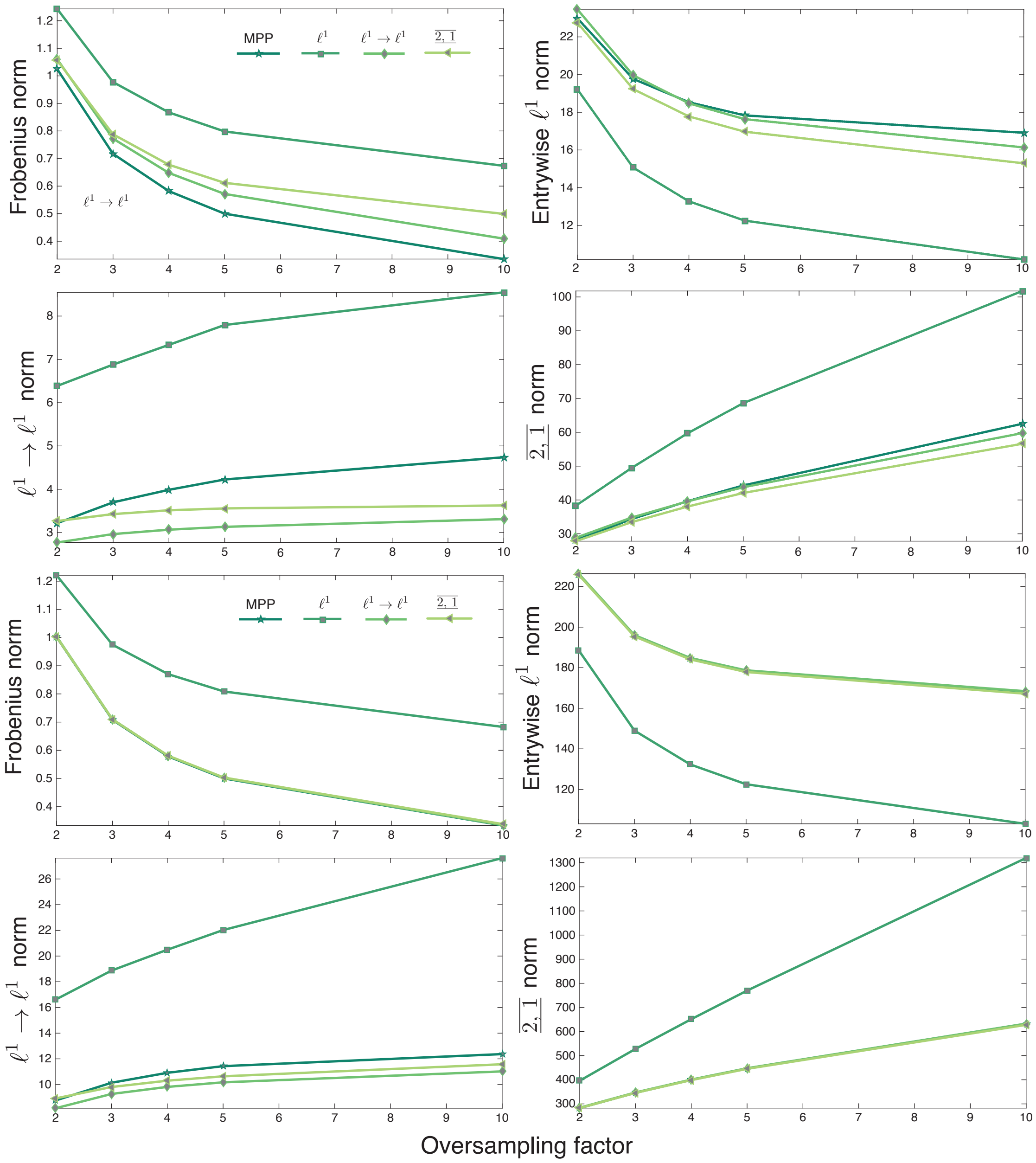


Figure 7.3: Average norms of different generalized inverses. Input matrices are drawn so that $a_{ij} \sim \mathcal{N}(0,1)$ independently with $m = 20$ rows (upper 4 graphs) and $m = 200$ rows (lower 4 graphs). The oversampling factor is defined as n/m where n is the number of columns.

7.5.2 Unbiasedness of Generalized Inverses

The MPP has certain “nice” properties that could be missing for other inverses. One such property is it being unbiased in a certain sense. That is, for $\boldsymbol{A}$ a random Gaussian matrix we

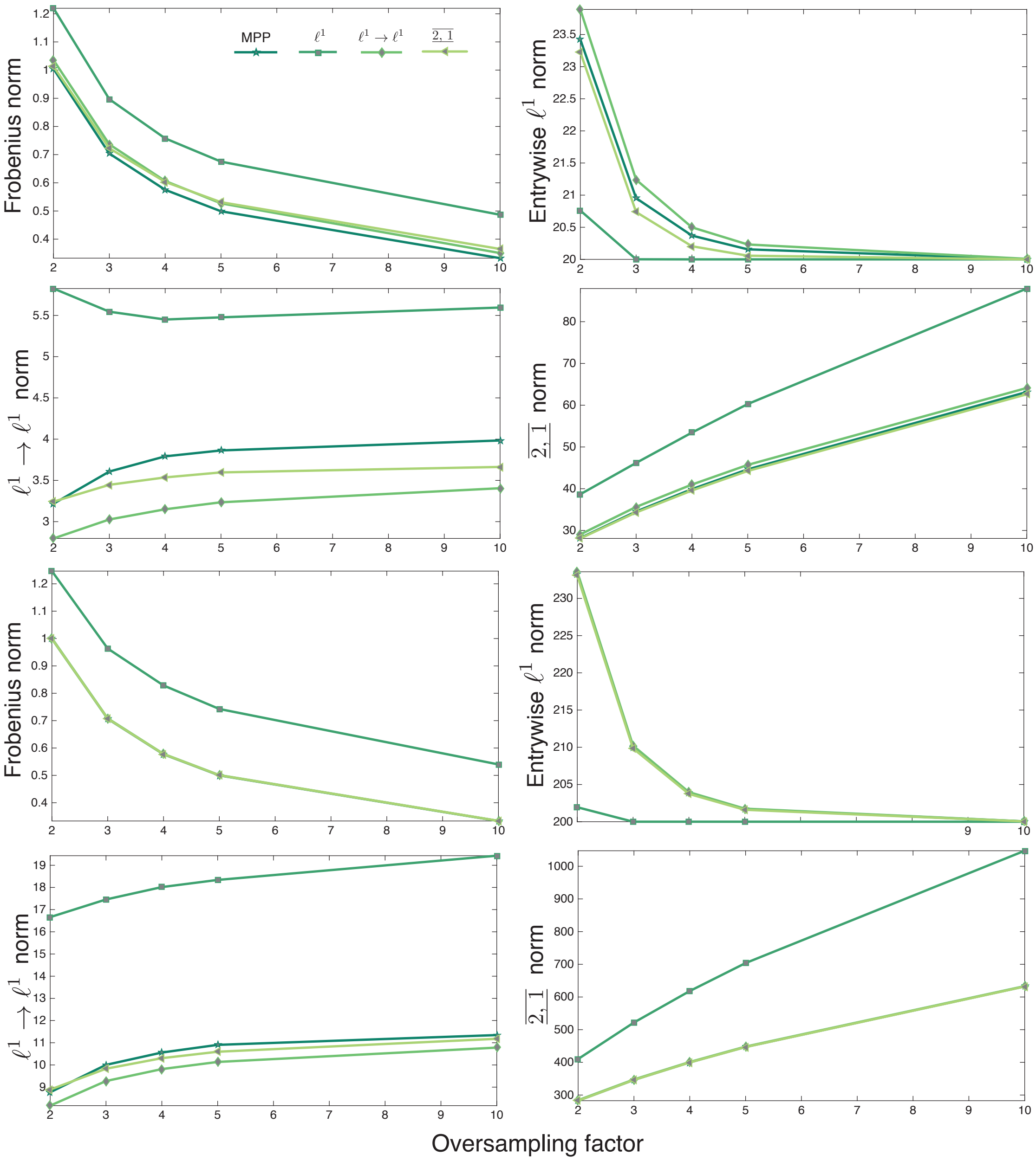


Figure 7.4: Average norms of different generalized inverses. Input matrices are drawn so that $a_{ij} \sim \text{Ber}(0.5)$ independently with $m = 20$ rows (upper 4 graphs) and $m = 200$ rows (lower 4 graphs). The oversampling factor is defined as n/m where n is the number of columns.

have that

$$\frac{n}{m}\mathbb{E}[\boldsymbol{A}^\dagger \boldsymbol{A}] = \boldsymbol{I}. \tag{7.46}$$

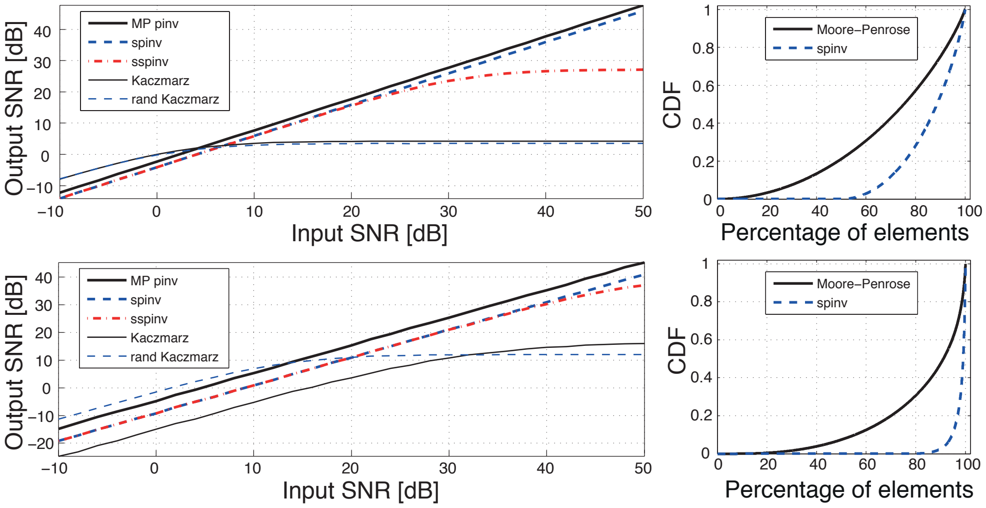


Figure 7.5: Output SNR and empirical cumulative distribution function (CDF) of matrix entry magnitudes for the MPP, $\operatorname{spinv}(\boldsymbol{A})$ and $\operatorname{sspinv}(\boldsymbol{A})$, for a full forward matrix (upper part) and a sparse forward matrix (lower part). A comparison is also shown with Kaczmarz and randomized Kaczmarz methods with a fixed number of scalar multiplication.

In other words, for this random matrix ensemble, applying $\boldsymbol{A}$ to a vector, and then the MPP to the measurements will on average retrieve the scaled version of the input vector. This property is useful in various iterative algorithms. To motivate it we consider the following generic procedure: let $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$, where $\boldsymbol{x}$ is the object we are interested in (*e.g.*, an image), $\boldsymbol{A}$ a dimensionality-reducing measurement system, and $\boldsymbol{y}$ the resulting measurements. One admissible estimate of $\boldsymbol{x}$ is given by $\boldsymbol{X}\boldsymbol{A}\boldsymbol{x}$, where $\boldsymbol{X} \in \mathcal{G}(\boldsymbol{A})$. If the dimensionality-reducing system is random, we can compute the expectation of the reconstructed vector as

$$\mathbb{E}[\boldsymbol{X}\boldsymbol{y}] = \mathbb{E}[\boldsymbol{X}\boldsymbol{A}\boldsymbol{x}] = \mathbb{E}[\boldsymbol{X}\boldsymbol{A}]\boldsymbol{x}. \tag{7.47}$$

Provided that $\mathbb{E}[\boldsymbol{X}\boldsymbol{A}] = \frac{m}{n}\boldsymbol{I}$, we will obtain, on average, a scaled version of the object we wish to reconstruct.[4]

Clearly, this property will not hold for generalized inverses obtained by inverting a particular minor of the input matrix. As we show next, it does hold for a large class of norm-minimizing generalized inverses.

---

[4] This linear step is usually part of a more complicated algorithm which also includes a nonlinear *denoising* step (*e.g.*, thresholding, non-local means). If the denoising step is a contraction in some sense (*i.e.* it brings us closer to the object we are reconstructing), the following scheme (or a similar one) will work well: $\boldsymbol{x}^{(k+1)} \overset{\text{def}}{=} \eta(\boldsymbol{x}^{(k)} + \frac{n}{m}\boldsymbol{X}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}^{(k)}))$.

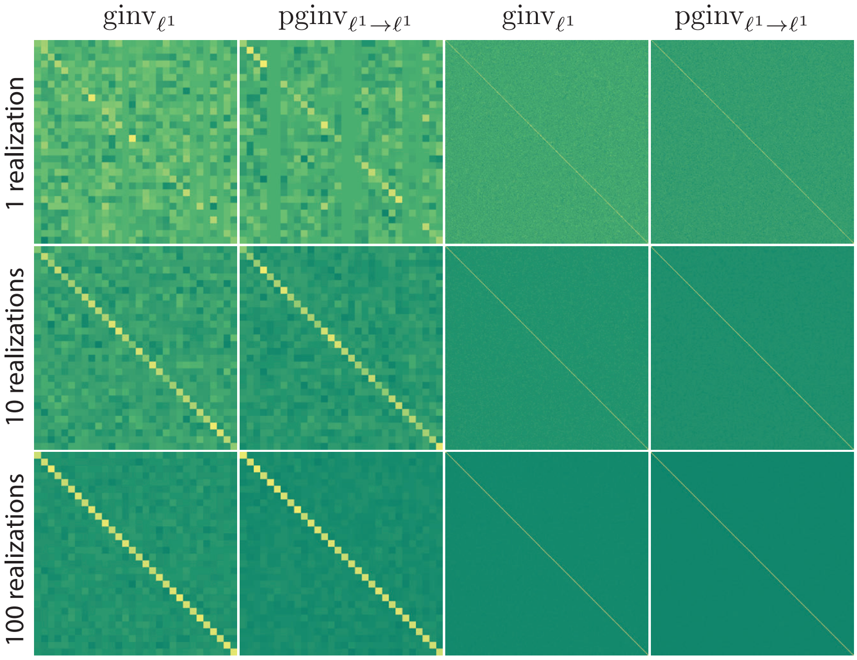


Figure 7.6: Illustration of the unbiasedness property. The left half shows $\mathrm{ginv}_{\ell^1}(\boldsymbol{A})\boldsymbol{A}$ averaged over the indicated number of realizations of $\boldsymbol{A} \in \mathbb{R}^{10\times 30}$ with a_{ij}. On the right, the same is shown for $\boldsymbol{A} \in \mathbb{R}^{100\times 300}$.

Theorem 7.5

Let $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $m < n$ *be a random matrix with iid columns such that* $a_{ij} \sim (-a_{ij})$. *Let further* $\|\cdot\|_\nu$ *be any matrix norm such that* $\|\boldsymbol{\Pi}\cdot\|_\nu = \|\cdot\|_\nu$ *and* $\|\boldsymbol{\Sigma}\cdot\|_\nu = \|\cdot\|_\nu$, *for any permutation matrix* $\boldsymbol{\Pi}$ *and modulation matrix* $\boldsymbol{\Sigma} = \mathrm{diag}(\boldsymbol{\sigma})$, *where* $\boldsymbol{\sigma} \in \{-1,1\}^n$. *Then, if* $\mathrm{ginv}_\nu(\boldsymbol{A})$ *and* $\mathrm{pginv}_\nu(\boldsymbol{A})$ *are singletons for all* $\boldsymbol{A}$, *we have*

$$\mathbb{E}[\mathrm{ginv}_\nu(\boldsymbol{A})\boldsymbol{A}] = \mathbb{E}[\mathrm{pginv}_\nu(\boldsymbol{A})\boldsymbol{A}] = \tfrac{m}{n}\boldsymbol{I}_n \tag{7.48}$$

More generally, consider a function $f : C \mapsto f(C) \in C \subset \mathbb{R}^{n\times m}$ *that selects a particular representative for any bounded convex set* C, *and assume that* $f(\boldsymbol{U}C) = \boldsymbol{U}f(C)$ *for any unitary matrix* $\boldsymbol{U}$ *and any* C. *Examples of such functions* f *include: selecting the centroid of the convex set, or selecting its element with minimum Frobenius norm. We have*

$$\mathbb{E}[f(\mathrm{ginv}_\nu(\boldsymbol{A}))\boldsymbol{A}] = \mathbb{E}[f(\mathrm{pginv}_\nu(\boldsymbol{A}))\boldsymbol{A}] = \tfrac{m}{n}\boldsymbol{I}_n. \tag{7.49}$$

Remark 4

This includes all classical norms (invariance to row permutations and sign changes) as well as any left unitarily invariant norm (permutations and sign changes are unitary).

To prove the theorem we use the following lemma,

Lemma 7.6

Let $\boldsymbol{U} \in \mathbb{C}^{n\times n}$ *be an invertible matrix, and* $\|\cdot\|_\nu$ *a norm such that* $\|\boldsymbol{U}\cdot\|_\nu = \|\cdot\|_\nu$. *Then the following claims hold,*

$$\operatorname{ginv}_\nu(\boldsymbol{A}\boldsymbol{U}) = \boldsymbol{U}^{-1}\operatorname{ginv}_\nu(\boldsymbol{A}), \tag{7.50}$$

$$\operatorname{pginv}_\nu(\boldsymbol{A}\boldsymbol{U}) = \boldsymbol{U}^{-1}\operatorname{pginv}_\nu(\boldsymbol{A}) \tag{7.51}$$

for any $\boldsymbol{A}$.

Proof of the lemma : We only prove the first claim; the remaining parts follow analogously using that $\operatorname{pginv}_\nu(\boldsymbol{A}) = \operatorname{ginv}_\mu(\boldsymbol{A})$ where $\|\cdot\|_\mu \overset{\text{def}}{=} \|\cdot\boldsymbol{A}\|_\nu = \|\boldsymbol{U}\cdot\boldsymbol{A}\|_\nu = \|\boldsymbol{U}\cdot\|_\mu$.

Feasibility: $(\boldsymbol{A}\boldsymbol{U})(\boldsymbol{U}^{-1}\operatorname{ginv}_\nu(\boldsymbol{A})) = \boldsymbol{A}\operatorname{ginv}_\nu(\boldsymbol{A}) = \boldsymbol{I}_m$.

Optimality: Consider any $\boldsymbol{X} \in \mathcal{G}(\boldsymbol{A}\boldsymbol{U})$. Since $(\boldsymbol{A}\boldsymbol{U})\boldsymbol{X} = \boldsymbol{I}_m = \boldsymbol{A}(\boldsymbol{U}\boldsymbol{X})$, the matrix $\boldsymbol{U}\boldsymbol{X}$ belongs to $\mathcal{G}(\boldsymbol{A})$ hence

$$\|\boldsymbol{X}\|_\nu = \|\boldsymbol{U}\boldsymbol{X}\|_\nu \geqslant \|\operatorname{ginv}_\nu(\boldsymbol{A})\|_\nu = \|\boldsymbol{U}\boldsymbol{U}^{-1}\operatorname{ginv}_\nu(\boldsymbol{A})\|_\nu = \|\boldsymbol{U}^{-1}\operatorname{ginv}_\nu(\boldsymbol{A})\|_\nu. \tag{7.52}$$

■

Proof (Theorem 7.5) : Since the matrix columns are iid, $\boldsymbol{A}$ is distributed identically to $\boldsymbol{A\Pi}$ for any permutation matrix $\boldsymbol{\Pi}$. This implies that functions of $\boldsymbol{A}$ and $\boldsymbol{A\Pi}$ have the same distribution(s). Thus the sets $f(\operatorname{ginv}_\nu(\boldsymbol{A}))\boldsymbol{A}$ and $f(\operatorname{ginv}_\nu(\boldsymbol{A\Pi}))\boldsymbol{A\Pi}$ are identically distributed. Using Lemma 7.6 with $\boldsymbol{U} = \boldsymbol{\Pi}$, we have that

$$\begin{aligned}\boldsymbol{M} \overset{\text{def}}{=} \mathbb{E}[f(\operatorname{ginv}_\nu(\boldsymbol{A}))\boldsymbol{A}] &= \mathbb{E}[f(\operatorname{ginv}_\nu(\boldsymbol{A\Pi}))\boldsymbol{A\Pi}] &(7.53)\\ &= \mathbb{E}[f(\boldsymbol{\Pi}^*\operatorname{ginv}_\nu(\boldsymbol{A}))\boldsymbol{A\Pi}] &(7.54)\\ &= \mathbb{E}[\boldsymbol{\Pi}^* f(\operatorname{ginv}_\nu(\boldsymbol{A}))\boldsymbol{A\Pi}] = \boldsymbol{\Pi}^*\boldsymbol{M}\boldsymbol{\Pi}. &(7.55)\end{aligned}$$

This is more explicitly written $m_{ij} = m_{\pi(i)\pi(j)}$ for all i, j and π the permutation associated to the permutation matrix $\boldsymbol{\Pi}$. Since this holds for any permutation matrix, we can write

$$\boldsymbol{M} = \begin{bmatrix} c & b & \cdots & b\\ b & c & \cdots & b\\ \vdots & & \ddots & \vdots\\ b & b & \cdots & c\end{bmatrix}, \tag{7.56}$$

We compute the value of $c = \frac{m}{n}$ as follows:

$$\begin{aligned} nc = \operatorname{trace}\ &\mathbb{E}[f(\operatorname{ginv}_\nu(\boldsymbol{A}))\boldsymbol{A}]\\ &= \mathbb{E}[\operatorname{trace}\ f(\operatorname{ginv}_\nu(\boldsymbol{A}))\boldsymbol{A}]\\ &= \mathbb{E}[\operatorname{trace}\ \boldsymbol{A}f(\operatorname{ginv}_\nu(\boldsymbol{A}))] \qquad (7.57)\\ &= \operatorname{trace}\ \boldsymbol{I}_m\\ &= m.\end{aligned}$$

To show that $b = 0$, we observe that since $a_{ij} \sim (-a_{ij})$, the matrices $\boldsymbol{A}$ and $\boldsymbol{A\Sigma}$ have the same distribution. As above, using Lemma 7.6 with $\boldsymbol{U} = \boldsymbol{\Sigma}$, this implies $\boldsymbol{M} = \boldsymbol{\Sigma M \Sigma}$ for any modulation matrix $\boldsymbol{\Sigma}$, that is to say $m_{ij} = \sigma_i m_{ij}\sigma_j$ for any i, j and $\boldsymbol{\sigma} \in \{-1, 1\}^n$. It follows that $m_{ij} = 0$ for $i \neq j$. Since we already established that $c_i \equiv \frac{m}{n}$ we conclude that $\mathbb{E}[f(\operatorname{ginv}_\nu(\boldsymbol{A}))\boldsymbol{A}] = \frac{m}{n}\boldsymbol{I}_n$. ■

The conditions of Theorem 7.5 are satisfied by many random matrices, *e.g.*, by iid Gaussian matrices.

7.5.3 Poor man's ℓ^p minimization revisited

As pointed out in the introduction, the generalized inverse $\mathrm{ginv}_{\ell^1 \to \ell^1}(\boldsymbol{A})$ is the one which minimizes the worst case blowup of the ℓ^1-norm between $\boldsymbol{z}$, the minimum ℓ^1 norm vector such that $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{z}$, and the linear estimate $\boldsymbol{X}\boldsymbol{y}$ where $\boldsymbol{X} \in \mathcal{G}(\boldsymbol{A})$. In this sense, $\mathrm{ginv}_{\ell^1 \to \ell^1}(\boldsymbol{A})$ provides the best worst-case poor man's (linear) ℓ^1 minimization, and solves

$$\inf_{\boldsymbol{X} \in \mathcal{G}(\boldsymbol{A})} \sup_{y \neq 0} \frac{\|\boldsymbol{X}\boldsymbol{y}\|_1}{\inf_{\boldsymbol{z}:\boldsymbol{A}\boldsymbol{z}=\boldsymbol{y}} \|\boldsymbol{z}\|_1}. \tag{7.58}$$

Instead of minimizing the worst-case blowup, we might want to minimize average-case ℓ^p blowup over a given class of input vectors. Let $\boldsymbol{u}$ be a random vector with the probability density function given by $f_{\boldsymbol{u}}$. Given $\boldsymbol{A}$, our goal is to minimize $\mathbb{E}[\|\boldsymbol{X}\boldsymbol{A}\boldsymbol{u}\|_p]$, which is the average ℓ^p norm blow-up raised to pth power. We replace this minimization by a simpler proxy: we aim to minimize $\mathbb{E}[\|\boldsymbol{X}\boldsymbol{A}\boldsymbol{u}\|_p^p]^{\frac{1}{p}}$. It is not difficult to verify that this expectation defines a norm (or a semi-norm, depending on $f_{\boldsymbol{u}}$).

Interestingly, for certain densities $f_{\boldsymbol{u}}$ this leads back to minimization of standard matrix norms:

Proposition 7.3

Assume that $\boldsymbol{u} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n)$. Then we have

$$\operatorname*{arg\,min}_{\boldsymbol{X} \in \mathcal{G}(\boldsymbol{A})} \left(\mathbb{E}[\|\boldsymbol{X}\boldsymbol{A}\boldsymbol{u}\|_p^p]\right)^{1/p} = \mathrm{pginv}_{\underline{\overline{2,p}}}(\boldsymbol{A}) \tag{7.59}$$

Remark 5

This result is intuitively pleasing. It is known that the $\underline{\overline{2,1}}$ mixed norm promotes row sparsity, thus the resulting $\boldsymbol{X}$ will have rows set to zero. Therefore, even if the result might have been predicted, it is interesting to see how a generic requirement to have a small ℓ^1 norm of the output leads to a known group sparsity cost function on the product matrix.

Proof:

$$\mathbb{E}_{\boldsymbol{u}}\big[\|\boldsymbol{X}\boldsymbol{A}\boldsymbol{u}\|_p^p\big]. = \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{u}}\big[|(\boldsymbol{X}\boldsymbol{A}\boldsymbol{u})_i|^p\big] \tag{7.60}$$

Because $\boldsymbol{u}$ is centered normal with covariance $\boldsymbol{I}_n$, the covariance matrix of $\boldsymbol{X}\boldsymbol{A}\boldsymbol{u}$ is $\boldsymbol{K} = (\boldsymbol{X}\boldsymbol{A})(\boldsymbol{X}\boldsymbol{A})^*$. Individual components are distributed according to $(\boldsymbol{X}\boldsymbol{A}\boldsymbol{u})_i \sim \mathcal{N}(0, \boldsymbol{K}_{ii}) = \mathcal{N}(0, \|\boldsymbol{x}^i\boldsymbol{A}\|_2^2)$. A straightforward computation shows that

$$\mathbb{E}\big[|(\boldsymbol{X}\boldsymbol{A}\boldsymbol{u})_i|^p\big] = \frac{2^{p/2}\,\Gamma\left(\frac{1+p}{2}\right)}{\sqrt{\pi}} \|\boldsymbol{x}^i\boldsymbol{A}\|_2^p. \tag{7.61}$$

We can then continue writing

$$\mathbb{E}\big[\|\boldsymbol{X}\boldsymbol{A}\boldsymbol{u}\|_p^p\big] = \sum_{i=1}^{n} \|\boldsymbol{x}^i\boldsymbol{A}\|_2^p = \|\boldsymbol{X}\boldsymbol{A}\|_{\underline{\overline{2,p}}}, \tag{7.62}$$

and the claim follows. ■

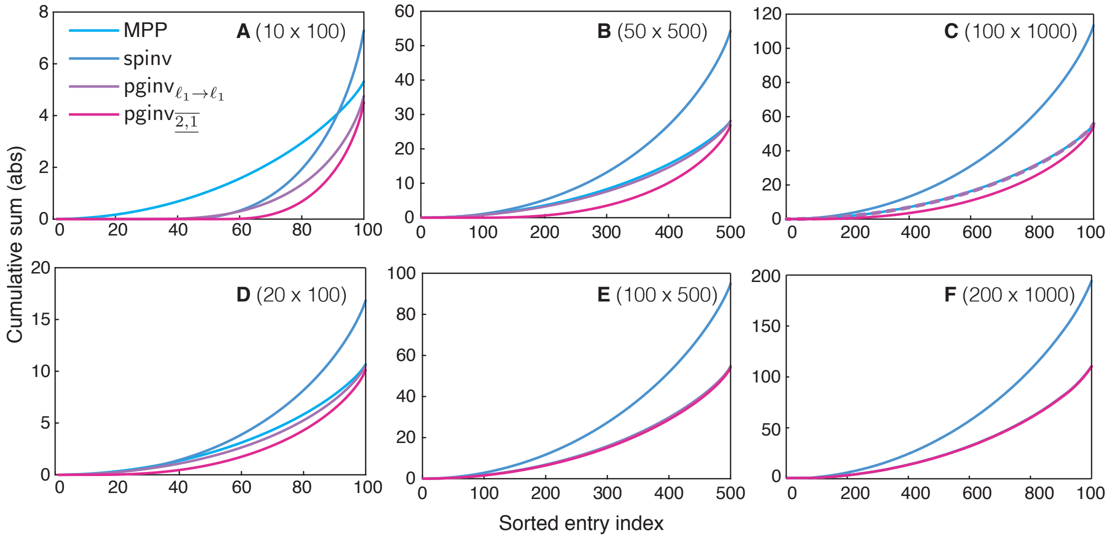


Figure 7.7: Cumulative sum profiles achieved by various generalized inverses for different matrix sizes. For every subfigure, 500 realizations of $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ were generated, with $a_{ij} \sim \mathcal{N}(0,1)$ iid. The dimensions m and n are indicated in parenthesis after the callout. The random matrix $\boldsymbol{A}$ was used to produce measurements $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$, where $\boldsymbol{x}$ was generated randomly for every realization of $\boldsymbol{A}$: the support of the $m/2$ non-zero entries was chosen uniformly at random, and the entries were again iid Gaussians. Four different *reconstructions* $\widehat{\boldsymbol{x}}$ were then obtained by four different generalized inverses, and for each of them we plotted the cumulative sum of the sorted entries of $|\widehat{\boldsymbol{x}}|$ (average over 500 realizations).

Numerical Experiment Figure 7.7 shows cumulative sums of sorted entry magnitudes of output vectors produced by different generalized inverses, for various problem sizes. We can see that $\mathrm{pginv}_{\underline{2,1}}(\boldsymbol{A})$ outputs more smaller elements, especially for smaller problem size. Nevertheless, the average ℓ^1-norm (the final value of the curve) is only slightly better than the one achieved by the MPP. Between the MPP and the $\mathrm{pginv}_{\underline{2,1}}(\boldsymbol{A})$ we have $\mathrm{pginv}_{\ell^1 \to \ell^1}(\boldsymbol{A})$. Interestingly, for larger problem sizes all inverses except the sparse pseudoinverse have identical profiles. This, together with the findings from Figure 7.3 and Figure 7.4 suggests that for large random matrices, all inverses become close to the MPP.

7.6 Matrices Having the Same Inverse for Many Norms

In the previous section, we saw that a large class of matrix norms leads to the Moore-Penrose pseudoinverse. In this section we discuss some classes of matrices whose generalized inverses minimize multiple norms.

7.6.1 Matrices which MPP has Entries on the Unit Circle

We start by an example where minimizing norms which usually do not give the MPP results in the MPP. When all entries of $\boldsymbol{A}^\dagger$ have the same magnitude, we can show that the MPP actually minimizes many norms beyond those already covered by Corollary 7.2.

Theorem 7.6

Let $\boldsymbol{A} \in \mathbb{C}^{m\times n}$. Suppose that all entries of $\boldsymbol{X} = \boldsymbol{A}^\dagger$ have the same magnitude, $|x_{ij}| = c$. Then the following statements are true,

(i) For $1 \leqslant p \leqslant \infty$, $0 < q \leqslant \infty$, we have

$$\boldsymbol{A}^\dagger \in \mathrm{ginv}_{|p,q|}(\boldsymbol{A}).$$

This set is a singleton for $1 < p < \infty$ and $0 < q < \infty$.

(ii) For $1 \leqslant p, q \leqslant \infty$, we have

$$\boldsymbol{A}^\dagger \in \mathrm{ginv}_{\overline{\underline{p,q}}}(\boldsymbol{A}).$$

This set is a singleton for $1 < p < q < \infty$.

Example 7.2

Primary examples of matrices $\boldsymbol{A}$ satisfying the assumptions of Theorem 7.6 are tight frames $\boldsymbol{A}$ with entries of constant magnitude, such as the partial Fourier matrix, $\boldsymbol{A} = \boldsymbol{RF}$, and the partial Hadamard matrix, $\boldsymbol{A} = \boldsymbol{RH}$, where $\boldsymbol{R}$ is the restriction of the identity matrix $\boldsymbol{I}_n$ to some arbitrary subset of m rows, and $\boldsymbol{F}$ (resp. $\boldsymbol{H}$) denote the Fourier (resp. a Hadamard) matrix of size n.

Indeed, when $\boldsymbol{A}$ is a tight frame, we have $\boldsymbol{AA}^* \propto \cdot \boldsymbol{I}_m$, hence $\boldsymbol{A}^\dagger = \boldsymbol{A}^*(\boldsymbol{AA}^*)^{-1} \propto \boldsymbol{A}^*$. When in addition the entries of $\boldsymbol{A}$ have equal magnitude, so must the entries of $\boldsymbol{A}^\dagger$.

Proof: Consider $\boldsymbol{E}$ a matrix and denote its columns as $\boldsymbol{\varepsilon}_j$. If the matrix $\boldsymbol{Y} = \boldsymbol{X} + \boldsymbol{E}$ still belongs to $\mathcal{G}(\boldsymbol{A})$, then for each column we have

$$\boldsymbol{A}(\boldsymbol{x}_j + \boldsymbol{\varepsilon}_j) = \boldsymbol{e}_j = \boldsymbol{A}\boldsymbol{x}_j. \tag{7.63}$$

That is to say $\boldsymbol{\varepsilon}_j$ must be in the nullspace of $\boldsymbol{A}$ for each column j. Since $\mathcal{N}(\boldsymbol{A}) = \mathrm{span}(\boldsymbol{A}^*)^\perp = \mathrm{span}(\boldsymbol{X})^\perp$, $\boldsymbol{\varepsilon}_j$ must be orthogonal to any column of $\boldsymbol{X}$, and in particular to $\boldsymbol{x}_j$. As a result,

$$\begin{aligned} \langle \boldsymbol{x}_j, \boldsymbol{\varepsilon}_j \rangle &= 0 \text{ for each column } j; & (7.64)\\ \text{and } \langle \boldsymbol{X}, \boldsymbol{E} \rangle &= 0 \text{ globally.} & (7.65) \end{aligned}$$

(i) To show this statement, it suffices to show that the columns $\boldsymbol{x}_j$ of $\boldsymbol{X} = \boldsymbol{A}^\dagger$ minimize all ℓ^p norms, $1 \leqslant p \leqslant \infty$, among the columns of $\boldsymbol{Y} \in \mathcal{G}(\boldsymbol{A})$.

By the convexity of $(\boldsymbol{\varepsilon}_R, \boldsymbol{\varepsilon}_I) \mapsto f(\boldsymbol{\varepsilon}_R, \boldsymbol{\varepsilon}_I) = \|\boldsymbol{x} + \boldsymbol{\varepsilon}_R + \mathrm{j}\boldsymbol{\varepsilon}_I)\|_p^p$, where the subscripts denote the real and imaginary parts of $\boldsymbol{x}$ and $\boldsymbol{\varepsilon}$ in a natural way and j is the imaginary unit, we have that $\|\boldsymbol{x} + \boldsymbol{\varepsilon}\|_p^p \geqslant \|\boldsymbol{x}\|_p^p + \langle \nabla f(\boldsymbol{0}, \boldsymbol{0}), (\boldsymbol{\varepsilon}_R, \boldsymbol{\varepsilon}_I)\rangle$. We can compute[5] that $\nabla f(\boldsymbol{\varepsilon}_R, \boldsymbol{\varepsilon}_I) = p\,|\boldsymbol{x}|^{(p-2)} \circ (\boldsymbol{x}_R, \boldsymbol{x}_I)$ so that for a column $\boldsymbol{x}_i$ of $\boldsymbol{X}$ with all entries of the same magnitude we have, using (7.64):

$$\begin{aligned} \|\boldsymbol{x}_i + \boldsymbol{\varepsilon}_i\|_p^p &\geqslant \|\boldsymbol{x}_i\|_p^p + p\left\langle |\boldsymbol{x}_i|^{\cdot(p-2)} \circ (\boldsymbol{x}_{iR}, \boldsymbol{x}_{iI}), (\boldsymbol{\varepsilon}_{iR}, \boldsymbol{\varepsilon}_{iI}) \right\rangle_{\mathbb{R}^{2n}} \\ &= \|\boldsymbol{x}_i\|_p^p + c(p, \boldsymbol{x}_i)\,\Re\,\langle \boldsymbol{x}_i, \boldsymbol{\varepsilon}_i \rangle_{\mathbb{C}^n} = \|\boldsymbol{x}_i\|_p^p. \end{aligned} \tag{7.66}$$

---

[5] $|\cdot|^{\cdot q}$ stands for the entrywise qth power of the magnitude of a vector or matrix; $\boldsymbol{x} \circ \boldsymbol{y}$ (resp. $\boldsymbol{X} \circ \boldsymbol{Y}$) is the entrywise multiplication.

Since this holds true for any $1 \leqslant p < \infty$, we also get $\|\boldsymbol{x}_j + \boldsymbol{\varepsilon}_j\|_\infty \geqslant \|\boldsymbol{x}_j\|_\infty$ by considering the limit when $p \to \infty$. For $1 < p < \infty$ and $0 < q < \infty$, the strict convexity of f and the strict monotonicity of the ℓ^q (quasi)norm imply that the inequality is strict whenever $\boldsymbol{E} \neq \boldsymbol{0}$, hence the uniqueness result.

(ii) We use the fact that $(\boldsymbol{E}_R, \boldsymbol{E}_I) \mapsto g(\boldsymbol{E}_R, \boldsymbol{E}_I) = \|\boldsymbol{X} + \boldsymbol{E}_R + \mathrm{j}\boldsymbol{E}_I\|_{\underline{\overline{p,q}}}^q$ is a convex function of $(\boldsymbol{E}_R, \boldsymbol{E}_I)$ (a composition of an affine function and a power ($q \geqslant 1$) of a (convex) norm). Therefore $\|\boldsymbol{X} + \boldsymbol{E}\|_{\underline{\overline{p,q}}}^q \geqslant \|\boldsymbol{X}\|_{\underline{\overline{p,q}}}^q + \langle \nabla g(\boldsymbol{0}, \boldsymbol{0}), (\boldsymbol{E}_R, \boldsymbol{E}_I)\rangle$. We can compute

$$\frac{\partial \|\boldsymbol{X}\|_{\underline{\overline{p,q}}}^q}{\partial x_{k\ell R}} = q\|\boldsymbol{x}^k\|_p^{q-p} |x_{k\ell}|^{p-2} x_{k\ell R}, \tag{7.67}$$

and a similar expression for the partial derivative with respect to the imaginary part $x_{k\ell I}$ so that

$$\nabla g(\boldsymbol{0}, \boldsymbol{0}) = q \operatorname{diag}(\|\boldsymbol{x}^1\|_p, \ldots, \|\boldsymbol{x}^n\|_p)^{q-p} |\boldsymbol{X}|^{\cdot(p-2)} \circ (\boldsymbol{X}_R, \boldsymbol{X}_I). \tag{7.68}$$

For $\boldsymbol{X}$ with all entries of the same magnitude we have, using (7.65):

$$\begin{aligned} \|\boldsymbol{X} + \boldsymbol{E}\|_{\underline{\overline{p,q}}}^q &\geqslant \|\boldsymbol{X}\|_{\underline{\overline{p,q}}}^q + c(p, q, \boldsymbol{X}) \, \langle (\boldsymbol{X}_R, \boldsymbol{X}_I), (\boldsymbol{E}_R, \boldsymbol{E}_I)\rangle \\ &= \|\boldsymbol{X}\|_{\underline{\overline{p,q}}}^q + c(p, q, \boldsymbol{X}) \, \Re\langle \boldsymbol{X}, \boldsymbol{E}\rangle = \|\boldsymbol{X}\|_{\underline{\overline{p,q}}}^q. \end{aligned} \tag{7.69}$$

As above the inequality $\|\boldsymbol{X} + \boldsymbol{E}\|_{\underline{\overline{p,q}}} \geqslant \|\boldsymbol{X}\|_{\underline{\overline{p,q}}}$ is extended to $p = \infty$ and/or $q = \infty$ by considering the limit. For $1 < p < \infty$ and $1 < q < \infty$ the function g is strictly convex implying that the inequality is strict whenever $\boldsymbol{E} \neq \boldsymbol{0}$, establishing the uniqueness result. ■

Example 7.3

Let $\boldsymbol{A} = [1,\ 1]$. *Then the MPP is given as*

$$\boldsymbol{A}^\dagger = \frac{1}{2}\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \tag{7.70}$$

and by Theorem 7.6 we know that $\|\boldsymbol{A}^\dagger\|_{|1,1|}$ *is optimal. Consider now the following matrix,*

$$\boldsymbol{A}^\ddagger(\alpha) = \begin{bmatrix} 1-\alpha \\ \alpha \end{bmatrix}, \tag{7.71}$$

for $0 \leqslant \alpha \leqslant 1$. *We have that* $\boldsymbol{A}\boldsymbol{A}^\ddagger = \boldsymbol{I}$, *so* $\boldsymbol{A}^\ddagger \in \mathcal{G}(\boldsymbol{A})$. *We also have that* $\|\boldsymbol{A}^\ddagger\|_{|1,1|} = \|\boldsymbol{A}^\dagger\|_{|1,1|} = 1$, *so the minimizer is not unique:* $\boldsymbol{A}^\ddagger(\alpha) \in \operatorname{ginv}_{|1,1|}(\boldsymbol{A}),\ \forall \alpha \in [0, 1]$.

Example 7.4

Similarly, let

$$\boldsymbol{A} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \end{bmatrix}, \tag{7.72}$$

and by Theorem 7.6 we know that $\|\boldsymbol{A}^\dagger\|_{|\infty,1|}$ *is optimal, because* $\boldsymbol{A}^\dagger = \frac{1}{4}\boldsymbol{A}^\top$. *But*

$$\boldsymbol{A}^\ddagger = \frac{1}{2}\begin{bmatrix} 1 & 0 \\ 0 & -1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}, \tag{7.73}$$

is also in $\mathcal{G}(\boldsymbol{A})$, *and* $\|\boldsymbol{A}^\ddagger\|_{\underline{\overline{1,\infty}}} = \|\boldsymbol{A}^\dagger\|_{\underline{\overline{1,\infty}}} = \frac{1}{2}$, *so again the minimizer is not unique.*

7.6.2 Matrices with a highly sparse generalized inverse

Next we consider matrices for which a single generalized inverse, which is *not* the MPP, simultaneously minimizes several norms. Consider $0 \leqslant p \leqslant 1$. It is known that if $\boldsymbol{x}$ is sufficiently sparse, then it can be uniquely recovered from $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$ by ℓ^p minimization. Denote $k_p(\boldsymbol{A})$ the largest integer k such that this holds true for any k-sparse vector $\boldsymbol{x}$:

$$k_p = \max \left\{ k \ : \ \operatorname*{arg\,min}_{\boldsymbol{A}\widehat{\boldsymbol{x}}=\boldsymbol{A}\boldsymbol{x}} \|\widehat{\boldsymbol{x}}\|_p = \{\boldsymbol{x}\}, \ \forall \boldsymbol{x} \text{ such that } \|\boldsymbol{x}\|_0 \leqslant k \right\}. \tag{7.74}$$

Using this definition we can then state the following theorem:

Theorem 7.7

Consider $\boldsymbol{A} \in \mathbb{C}^{m\times n}$, and assume there exists a generalized inverse $\boldsymbol{X} \in \mathcal{G}(\boldsymbol{A})$ such that every of its columns $\boldsymbol{x}_i$ is $k_{p_0}(\boldsymbol{A})$-sparse, for some $0 < p_0 \leqslant 1$. Then, for all $0 < q < \infty$, and all $0 \leqslant p \leqslant p_0$, we have

$$\operatorname{ginv}_{|p,q|}(\boldsymbol{A}) = \{\boldsymbol{X}\}.$$

Example 7.5

Let $\boldsymbol{A}$ be the Dirac-Fourier (resp. the Dirac-Hadamard) dictionary, $\boldsymbol{A} = [\boldsymbol{I}, \ \boldsymbol{F}]$ (resp. $\boldsymbol{A} = [\boldsymbol{I}, \ \boldsymbol{H}]$). It is known that $k_1(\boldsymbol{A}) \geqslant (1+\sqrt{m})/2$ (see, e.g., *[86]). Moreoever*

$$\boldsymbol{X} \overset{\text{def}}{=} \begin{bmatrix} \boldsymbol{I} \\ \boldsymbol{0} \end{bmatrix} \in \mathcal{G}(\boldsymbol{A})$$

has $k = 1$-sparse columns. As a result, $\operatorname{ginv}_{|p,q|}(\boldsymbol{A}) = \{\boldsymbol{X}\}$ for any $0 \leqslant p \leqslant 1$ and $0 < q < \infty$. On this specific case, the equality can also be checked for $q = \infty$. Thus, $\operatorname{ginv}_{|p,q|}(\boldsymbol{A})$ is distinct from the MPP of $\boldsymbol{A}$, $\boldsymbol{A}^\dagger = \boldsymbol{A}^*/2$.

Proof: It is known [86] that for $0 \leqslant p \leqslant p_0 \leqslant 1$ and any $\boldsymbol{A}$ we have $k_p(\boldsymbol{A}) \geqslant k_{p_0}(\boldsymbol{A})$. Hence, the column $\boldsymbol{x}_i$ of $\boldsymbol{X}$ is the unique minimum ℓ^p norm solution to $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{e}_i$. When $q < \infty$ this implies $\operatorname{ginv}_{|p,q|}(\boldsymbol{A}) = \{\boldsymbol{X}\}$. When $q = \infty$ this simply implies $\operatorname{ginv}_{|p,q|}(\boldsymbol{A}) \ni \boldsymbol{X}$. ■

7.7 Computation of Norm-Minimizing Generalized Inverses

In this section we discuss the computation of the generalized inverses associated with matrix norms. Interior point methods, and general-purpose solvers such as CVX are generally quite slow, especially for problems that cannot be reduced to vector problems. For the kind of problems that we need to solve, it appears more appropriate to use methods such as the projected gradient method, or the alternating direction method of multipliers (ADMM) [164], also known as the Douglas-Rachford splitting method [68].

To implement ADMM, we should know how to compute the proximal operators associated with matrix norms, or equivalently, projections onto various matrix norm balls. The attractive property is that we do not care about possible non-smoothness or non-strict convexity. On the other hand, if the corresponding proximal operators are difficult to compute, we may use the projected subgradient method that can handle non-smooth norms. We find ADMM more attractive because it yields to an immediate extension from ginv to pginv via the so called *linearized ADMM*. Thus we may use almost the same algorithm to optimize for $\|\boldsymbol{X}\|$ and for $\|\boldsymbol{X}\boldsymbol{A}\|$.

7.7.1 Norms that Reduce to the Vector Case

In certain cases, it is possible to reduce the computation of the norm-minimizing generalized inverse to a collection of independent vector problems. This is the case with entrywise norms and with columnwise mixed norms. As the entrywise norms are a special case of the columnwise mixed norms, we discuss the latter case.

Consider the minimization for $\operatorname{ginv}_{|p,q|}(\boldsymbol{A})$, $q < \infty$. It is straightforward to verify the following series of equivalences:

$$\min_{\boldsymbol{X} \in \mathcal{G}(\boldsymbol{A})} \left(\sum_i \|\boldsymbol{x}_i\|_p^q \right)^{\frac{1}{q}} \Leftrightarrow \min_{\boldsymbol{A}\boldsymbol{X}=\boldsymbol{I}} \sum_i \|\boldsymbol{x}_i\|_p^q \Leftrightarrow \left\{ \min_{\boldsymbol{A}\boldsymbol{x}_i = \boldsymbol{e}_i} \|\boldsymbol{x}_i\|_p, \ 1 \leqslant i \leqslant n \right\}. \tag{7.75}$$

We can therefore use our favorite algorithm for finding the ℓ^p-minimal solution to an underdetermined system of linear equations. The most interesting cases are probably for $p \in \{1, 2, \infty\}$. For $p = 2$, we of course get the MPP, and for the other two cases, we have efficient algorithms at our disposal.

7.7.2 ADMM for Norms that do not (Easily) Reduce to the Vector Case

Even for entrywise and columnwise mixed norms, things get more complicated when instead of $\operatorname{ginv}_{|p,q|}(\boldsymbol{A})$ we try to compute $\operatorname{pginv}_{|p,q|}(\boldsymbol{A})$. This is because the objective $\|\boldsymbol{X}\boldsymbol{A}\|_{|p,q|}$ now mixes the columns of $\boldsymbol{X}$ so that they cannot be untangled. Similar issues arise when trying to compute $\operatorname{ginv}_{\underline{\overline{p,q}}}(\boldsymbol{A})$. We will see that this issue is elegantly adressed by the linearized ADMM.

7.7.3 Alternating Direction Method of Multipliers

ADMM is a method to solve problems of the form

$$\text{minimize} \ \ f(x) + g(x), \tag{7.76}$$

where in our case x is a matrix $\boldsymbol{X}$, and g is a monotonically increasing function of a matrix norm $\|\boldsymbol{X}\|$. The reason why we choose g and not f to be the matrix norm has to do with efficient computation of pginv(), as discussed below. ADMM relies on the iterative computation of proximal operators

$$\operatorname{prox}_\varphi(y) = \arg\min_z \tfrac{1}{2}\|y - x\|_2^2 + \varphi(x).$$

Proximal operator of the indicator function $I_{\mathcal{G}(\boldsymbol{A})}(\boldsymbol{X})$.

To cast our minimization in the form (7.76), we first rewrite the constraint using the indicator function. The indicator function of a set $\mathcal{S}$ is defined as

$$I_{\mathcal{S}}(\boldsymbol{X}) \overset{\text{def}}{=} \begin{cases} 0, & \boldsymbol{X} \in \mathcal{S} \\ +\infty, & \boldsymbol{X} \notin \mathcal{S}. \end{cases} \tag{7.77}$$

Its proximal operator is just the Euclidean projection onto the set. In the case of the affine subspace $\mathcal{S} = \mathcal{G}(\boldsymbol{A})$, we can verify that this is given as

$$\operatorname{prox}_{I_{\mathcal{G}(\boldsymbol{A})}}(\boldsymbol{V}) = \operatorname{proj}_{\mathcal{G}(\boldsymbol{A})}(\boldsymbol{V}) = \boldsymbol{A}^\dagger + \boldsymbol{N}\boldsymbol{N}^*\boldsymbol{V}, \tag{7.78}$$

where the columns of $\boldsymbol{N}$ form an orthogonal basis for the nullspace of $\boldsymbol{A}$.

Proximal operators of matrix norms

Computing the proximal operator for various norms can get more tricky. We expound here the proximal operators for some mixed and some induced norms; the first ingredient is an expression for the proximal operator of a norm in terms of the projection onto the norm ball of the dual norm. We have that for any scalar $\lambda > 0$,

$$\operatorname{prox}_{\lambda\|\cdot\|}(\boldsymbol{V}) = \boldsymbol{V} - \lambda \operatorname{proj}_{\{\boldsymbol{X}:\|\boldsymbol{X}\|^* \leqslant 1\}}(\boldsymbol{V}) \tag{7.79}$$

with $\|\cdot\|^\star$ the dual norm to $\|\cdot\|$. This means that we can compute the proximal operator efficiently if we can project efficiently, and vice-versa.

We continue with a useful lemma:

Lemma 7.7

The dual norm of $\|\cdot\|_{\ell^1\to\ell^q} = \|\cdot\|_{|q,\infty|}$ *is*

$$\|\cdot\|^*_{\ell^1\to\ell^q} = \|\cdot\|_{|q^*,1|},$$

and the dual norm of $\|\cdot\|_{\ell^p\to\ell^\infty} = \|\cdot\|_{\underline{\overline{p^*,\infty}}}$ *is*

$$\|\cdot\|^*_{\ell^p\to\ell^\infty} = \|\cdot\|_{\underline{\overline{p,1}}},$$

where $1/p + 1/p^* = 1/q + 1/q^* = 1$.

Proof: The first result is a direct consequence of the characterization (7.27) of $\ell^1 \to \ell^q$ norms in terms of columnwise mixed norms $|q,\infty|$, and of [192, Lemma 3]. The second result follows from the characterization (7.33) of $\ell^p \to \ell^\infty$ norms in terms of $\underline{\overline{p^*,\infty}}$ and, again, of [192, Lemma 3]. ■

Combining (7.79) with Lemma 7.7 shows that for $\|\cdot\|_{\ell^1\to\ell^q}$, where $q \in \{1,2,\infty\}$, computing the proximal operator means projecting onto the ball of the $\|\cdot\|_{|q^*,1|}$ norm. The good news is that these projections can be computed efficiently [192, 172]. Even for a general q, the projection can be computed efficiently, but the algorithm becomes more involved [125, 217]. In summary, we can efficiently compute the proximal mapping for induced norms $\|\cdot\|_{\ell^1\to\ell^q}$, and equivalently for induced norms $\|\cdot\|_{\ell^p\to\ell^\infty}$, because these read as proximal mappings for certain mixed norms.

As far as the columnwise mixed norms go, we do not really have to compute the proximal mapping for the norm itself, but rather for $\|\cdot\|^p_{|p,q|}$. This can be again achieved by decoupling because

$$\begin{aligned}
\operatorname{prox}_{\lambda\|\cdot\|^p_{|p,q|}}(\boldsymbol{V}) &= \operatorname*{arg\,min}_{\boldsymbol{X}} \sum_i \|\boldsymbol{x}_i\|_q^p + \frac{1}{2\lambda}\|\boldsymbol{X} - \boldsymbol{V}\|_F^2 \\
&= \operatorname*{arg\,min}_{\boldsymbol{X}} \sum_i \left(\|\boldsymbol{x}_i\|_q^p + \frac{1}{2\lambda}\|\boldsymbol{x}_i - \boldsymbol{v}_i\|_2^2 \right).
\end{aligned} \tag{7.80}$$

We now converted the problem to computing $\operatorname{prox}_{\lambda\|\cdot\|_q^p}$, and the task is to find efficient algorithms to compute it. We focus again on the interesting (and simpler) case of $p = 1$. Here, we can use known proximal operators for ℓ^q norms. Alternatively, we could directly compute the proximal mapping from known projection algorithms and using root finding.

Finally, we note that an analogous argumentation can be made for rowwise norms: Their proximal mappings can be constructed from vector proximal mappings as well.

Our minimization program now becomes

$$\text{minimize } h(\|\boldsymbol{X}\|) + I_{\mathcal{G}(\boldsymbol{A})}(\boldsymbol{X}). \tag{7.81}$$

To apply ADMM, we need to compute the proximal operators of functions f and g.

7.7.4 Generic ADMM for matrix norm minimization

The generic ADMM updates for computing $\operatorname{ginv}_\nu(\boldsymbol{A})$ can be written as [164],

$$\begin{aligned}
\boldsymbol{X}^{k+1} &\overset{\text{def}}{=} \boldsymbol{A}^\dagger + \boldsymbol{N}\boldsymbol{N}^*(\boldsymbol{X}^k - \boldsymbol{U}^k) \\
\boldsymbol{Z}^{k+1} &\overset{\text{def}}{=} \operatorname{prox}_{\lambda\|\cdot\|_\nu}(\boldsymbol{X}^{k+1} + \boldsymbol{U}^k) \\
\boldsymbol{U}^{k+1} &\overset{\text{def}}{=} \boldsymbol{U}^k + \boldsymbol{X}^{k+1} - \boldsymbol{Z}^{k+1}.
\end{aligned} \tag{7.82}$$

While there are a number of references that study the choice of the regularization parameter λ for particular f and g, this choice still remains something of a black art. A discussion of this topic falls out of the scope of our work. In our implementations we used values for which the algorithm was empirically verified to converge.

The convenience of ADMM is that we can easily adapt this for norms on the matrix $\boldsymbol{X}\boldsymbol{A}$, without computing the new proximal operator. In particular, the ADMM solves the following problem,

$$\text{minimize } f(\boldsymbol{x}) \;+\; g(\boldsymbol{A}\boldsymbol{x}), \tag{7.83}$$

with updates involving $\boldsymbol{A}$ and $\boldsymbol{A}^*$. If we think of this as applying a linear operator and its adjoint, we can adapt it to our problem which has the form

$$\text{minimize } f(\boldsymbol{X}) \;+\; g(\boldsymbol{X}\boldsymbol{A}), \tag{7.84}$$

where we consider g to be the norm that we aim to minimize. It is easy to verify that the adjoint of postmultiplying an $n \times m$ matrix by $\boldsymbol{A}$ is simply postmultiplication by $\boldsymbol{A}^*$,

$$\langle \boldsymbol{X}\boldsymbol{A}, \boldsymbol{Y} \rangle = \operatorname{trace}((\boldsymbol{X}\boldsymbol{A})^*\boldsymbol{Y}) = \operatorname{trace}(\boldsymbol{A}^*\boldsymbol{X}^*\boldsymbol{Y}) = \operatorname{trace}(\boldsymbol{X}^*\boldsymbol{Y}\boldsymbol{A}^*) = \left\langle \boldsymbol{X}, \boldsymbol{Y}\boldsymbol{A}^* \right\rangle. \tag{7.85}$$

The updates for $\operatorname{pginv}_{\|\cdot\|}(\boldsymbol{A})$ are then

$$\begin{aligned}
\boldsymbol{X}^{k+1} &\overset{\text{def}}{=} \boldsymbol{A}^\dagger + \boldsymbol{N}\boldsymbol{N}^* \Big[\boldsymbol{X}^k - (\mu/\lambda)(\boldsymbol{X}^k\boldsymbol{A} - \boldsymbol{Z}^k + \boldsymbol{U}^k)\boldsymbol{A}^*\Big] \\
\boldsymbol{Z}^{k+1} &\overset{\text{def}}{=} \operatorname{prox}_{\lambda\|\cdot\|_\nu}(\boldsymbol{X}^{k+1}\boldsymbol{A} + \boldsymbol{U}^k) \\
\boldsymbol{U}^{k+1} &\overset{\text{def}}{=} \boldsymbol{U}^k + \boldsymbol{X}^{k+1}\boldsymbol{A} - \boldsymbol{Z}^{k+1},
\end{aligned} \tag{7.86}$$

where $0 < \mu \leqslant \lambda/\|A\|^2$.

7.8 Conclusion

Our motivation in the quest for interesting norm-minimizing generalized inverses is twofold: 1) need for a *fast* generalized inverse, and 2) prospect of having operators that yield as poor man's ℓ^p minimizers. Often our developments seem Sisyphean, in that for various norms and matrices we just recover the optimality of the MPP. But there are cases such as the ℓ^1 entrywise norm, which always yield a different solution with interesting properties. Furthermore, it is valuable to know that in terms of many interesting optimality measures, there is no need to look beyond the MPP if we insist on linearity.

7.A Proof of Corollary 7.2

We begin by showing that the MPP is *a* minimizer of the considered norms. The result for $\|\cdot\|_{S_p}$ with $1 \leqslant p \leqslant \infty$, for the induced norm $\|\cdot\|_{\ell^2 \to \ell^2}$ and columnwise mixed norm $\|\cdot\|_{|2,2|}$ follow from their unitary invariance and Theorem 7.1. In contrast, the norms $\|\cdot \boldsymbol{A}\|_{S_p}$ generally fail to be unitarily invariant. Similarly, for $p \neq 2$ and $q \neq 2$, induced norms and columnwise mixed norms are *not* unitarily invariant, hence the results do not directly follow from Theorem 7.1. Instead, the reader can easily check that all considered norms are left unitarily invariant, hence the MPP is *a* minimizer by Theorem 7.2 and Corollary 7.1.

We now turn to uniqueness for Schatten norms and columnwise mixed norms. Since Schatten norms with $1 \leqslant p < \infty$ are fully unitarily invariant and associated to a strictly monotonic symmetric gauge function, the uniqueness result of Theorem 7.1 applies. To establish uniqueness with the norms $\|\cdot \boldsymbol{A}\|_{S_p}$, $1 \leqslant p < \infty$ we exploit the following useful Lemma (see, *e.g.*, [224, Lemma 7] and references therein for a proof).

Lemma 7.8

Let $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{C}^{m \times n}$ *be given as*

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{K}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}, \qquad \boldsymbol{Y} = \begin{bmatrix} \boldsymbol{K}_1 & \boldsymbol{K}_2 \\ \boldsymbol{K}_3 & \boldsymbol{K}_4 \end{bmatrix} \tag{7.87}$$

where the block $\boldsymbol{K}_1$ *is of size* $r \times r$*. Then we have that* $\sigma_j(\boldsymbol{X}) \leqslant \sigma_j(\boldsymbol{Y})$ *for* $1 \leqslant j \leqslant r$*, and for any unitarily invariant norm* $\|\cdot\|_\phi$ *associated to a symmetric gauge* ϕ*, we have* $\|\boldsymbol{X}\|_\phi \leqslant \|\boldsymbol{Y}\|_\phi$*. When* ϕ *is strictly monotonic,* $\|\boldsymbol{X}\|_\phi = \|\boldsymbol{Y}\|_\phi$ *if and only if* $\boldsymbol{K}_2, \boldsymbol{K}_3$ *and* $\boldsymbol{K}_4$ *are zero blocks.*

Considering $\boldsymbol{X} \in \mathcal{G}(\boldsymbol{A})$ and its representation as given in (7.15), using the unitary invariance of the Schatten norm, we have

$$\|\boldsymbol{X}\boldsymbol{A}\|_{S_p} = \|\boldsymbol{M}\boldsymbol{\Sigma}\|_{S_p} = \left\| \begin{bmatrix} \boldsymbol{I}_m & \boldsymbol{0} \\ \boldsymbol{S}\boldsymbol{\Sigma}_\square & \boldsymbol{0} \end{bmatrix} \right\|_{S_p} > \left\| \begin{bmatrix} \boldsymbol{I}_m & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \right\|_{S_p} = \|\boldsymbol{A}^\dagger \boldsymbol{A}\|_{S_p} \tag{7.88}$$

as soon as $\boldsymbol{S}\boldsymbol{\Sigma}_\square \neq \boldsymbol{0}$, that is to say whenever $\boldsymbol{S} \neq \boldsymbol{0}$. For both types of columnwise mixed norms with $1 \leqslant q < \infty$, the strictness of the inequality in Lemma 7.3 when $\boldsymbol{K}_2 \neq \boldsymbol{0}$ is easy to check, and we can apply the uniqueness result of Corollary 7.1.

Finally, we consider the possible lack of uniqueness. The construction of $\boldsymbol{A}^\ddagger \neq \boldsymbol{A}^\dagger$ in Example 7.1 provides a matrix $\boldsymbol{A}$ for which $\operatorname{ginv}_{S_\infty}(\boldsymbol{A}) \supset \{\boldsymbol{A}^\dagger, \boldsymbol{A}^\ddagger\}$ is not reduced to the MPP.

To prove uniqueness for $\operatorname{pginv}_{S_\infty}(\boldsymbol{A})$, consider again equation (7.88) with $p = \infty$. Denote $\boldsymbol{Z} \overset{\text{def}}{=} \boldsymbol{S}\boldsymbol{\Sigma}_\square \neq \boldsymbol{0}$, and let $\boldsymbol{z}^*$ be its first nonzero row. We bound the largest singular value σ_1 of the matrix $\boldsymbol{Y}$ defined as follows

$$\boldsymbol{Y} \overset{\text{def}}{=} \begin{bmatrix} \boldsymbol{I}_m & \boldsymbol{0} \\ \boldsymbol{z}^* & 0 \end{bmatrix}.$$

By the variational characterization of eigenvalues of a Hermitian matrix, we have

$$\sigma_1^2 = \max_{\|\boldsymbol{x}\|_2 = 1} \boldsymbol{x}^* \boldsymbol{Y} \boldsymbol{Y}^* \boldsymbol{x} = \max_{\|\boldsymbol{x}\|_2 = 1} \boldsymbol{x}^* \begin{bmatrix} \boldsymbol{I}_m & \boldsymbol{z} \\ \boldsymbol{z}^* & \|\boldsymbol{z}\|_2^2 \end{bmatrix} \boldsymbol{x}. \tag{7.89}$$

Partitioning $\boldsymbol{x}$ as $\boldsymbol{x}^* = (\tilde{\boldsymbol{x}}^*, \ x_n^*)$ with $x_n \in \mathbb{C}$ we write the maximization as

$$\max_{\|\tilde{\boldsymbol{x}}\|_2^2 + |x_n|^2 = 1} \|\tilde{\boldsymbol{x}}\|_2^2 + 2\Re(x_n^* \boldsymbol{z}^* \tilde{\boldsymbol{x}}) + |x_n|^2 \|\boldsymbol{z}\|_2^2. \tag{7.90}$$

Since $\boldsymbol{z} \neq 0$ by assumption, it must have at least one non-zero entry. Let i be the index of a non-zero entry, and restrict $\widetilde{\boldsymbol{x}}$ to the form $\widetilde{\boldsymbol{x}} = \alpha \boldsymbol{e}_i$, $\alpha \in \mathbb{C}$. We have that

$$\max_{\|\widetilde{\boldsymbol{x}}\|_2^2 + |x_n|^2 = 1} \|\widetilde{\boldsymbol{x}}\|_2^2 + 2\Re(x_n^* \boldsymbol{z}^* \widetilde{\boldsymbol{x}}) + |x_n|^2 \|\boldsymbol{z}\|_2^2 \geqslant \max_{|\alpha| \leqslant 1} |\alpha|^2 + 2\sqrt{1 - |\alpha|^2} \cdot |z_i| \cdot |\alpha| + (1 - |\alpha|^2)|z_i|^2, \tag{7.91}$$

where we used the fact that maximization over a smaller set can only diminish the optimum, as well as that $\|\boldsymbol{z}\|_2^2 \geqslant |z_i|^2$. Straightforward calculus shows that the maximum of the last expression is $1 + |z_i|^2$, which is strictly larger than 1. Since $\boldsymbol{Y}\boldsymbol{Y}^*$ is a principal minor of

$$\begin{bmatrix} \boldsymbol{I}_m & \boldsymbol{0} \\ \boldsymbol{S}\boldsymbol{\Sigma}_{\square} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{I}_m & \boldsymbol{0} \\ \boldsymbol{S}\boldsymbol{\Sigma}_{\square} & \boldsymbol{0} \end{bmatrix}^*,$$

it follows by Cauchy's interlacing theorem (or it could be seen directly from (7.89)) that

$$\left\| \begin{bmatrix} \boldsymbol{I}_m & \boldsymbol{0} \\ \boldsymbol{S}\boldsymbol{\Sigma}_{\square} & \boldsymbol{0} \end{bmatrix} \right\|_{S_\infty} \geqslant \sigma_1 \geqslant \sqrt{1 + |z_i|^2} \overset{\text{strictly}}{>} 1 = \|\boldsymbol{A}^\dagger \boldsymbol{A}\|_{S_\infty}.$$

A counterexample for $\operatorname{ginv}_{|2,\infty|}(\cdot)$ is as follows: consider $0 < \eta < 1$ and the following matrix,

$$\boldsymbol{A} = \begin{bmatrix} \eta^{-1} & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}. \tag{7.92}$$

Its MPP is simply $\boldsymbol{A}^\dagger = \begin{bmatrix} \eta & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}^\top$, and we have that $\|\boldsymbol{A}^\dagger\|_{|2,\infty|} = 1$. Among many other generalized inverses, one class is given as

$$\boldsymbol{A}^\ddagger(\alpha) = \begin{bmatrix} \eta & 0 \\ 0 & 1 \\ \alpha & 0 \end{bmatrix}. \tag{7.93}$$

Clearly for all α with $|\alpha| \leqslant \sqrt{1 - \eta^2}$, $\|\boldsymbol{A}^\ddagger(\alpha)\|_{|2,\infty|} = \|\boldsymbol{A}^\dagger\|_{|2,\infty|} = 1$, hence $\operatorname{ginv}_{|2,\infty|}(\cdot)$ is not a singleton. To construct a counterexample for $\operatorname{pginv}_{|2,\infty|}()$, consider a full rank matrix $\boldsymbol{A}$ with the SVD $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$. Then all the generalized inverses $\boldsymbol{A}^\ddagger$ are given as (7.15), so that

$$\|\boldsymbol{A}^\ddagger \boldsymbol{A}\|_{|2,\infty|} = \left\| \begin{bmatrix} \boldsymbol{V}_1^* \\ \boldsymbol{S}\boldsymbol{\Sigma}\boldsymbol{V}_1^* \end{bmatrix} \right\|_{|2,\infty|}, \tag{7.94}$$

where we applied the left unitary invariance and we partitioned $\boldsymbol{V} = [\,\boldsymbol{V}_1 \ \boldsymbol{V}_2\,]$. Clearly, setting $\boldsymbol{S} = \boldsymbol{0}$ gives the MPP and optimizes to norm. To construct a counterexample, note that generically $\boldsymbol{V}_1^*$ will have columns with different 2-norms. Let $\boldsymbol{v}$ be its column with the largest 2-norm; we choose a matrix $\boldsymbol{P}$ so that $\boldsymbol{P}\boldsymbol{\Sigma}\boldsymbol{v} = \boldsymbol{0}$ while $\boldsymbol{P}\boldsymbol{\Sigma}\boldsymbol{V}_1^* \neq \boldsymbol{0}$. Then there exists ε_0 so that for any $\varepsilon \leqslant \varepsilon_0$,

$$\left\| \begin{bmatrix} \boldsymbol{V}_1^* \\ \varepsilon \boldsymbol{P}\boldsymbol{\Sigma}\boldsymbol{V}_1^* \end{bmatrix} \right\|_{|2,\infty|} = \left\| \begin{bmatrix} \boldsymbol{V}_1^* \\ \boldsymbol{0} \end{bmatrix} \right\|_{|2,\infty|} = \|\boldsymbol{A}^\dagger \boldsymbol{A}\|_{|2,\infty|}. \tag{7.95}$$

We can reuse counterexample (7.92) to show the lack of uniqueness for the induced norms (note that we already have two counterexamples: for $p = 1$ as this gives back the columnwise norm, and for $p = 2$ as this gives back the spectral norm which is the Schatten infinity norm).

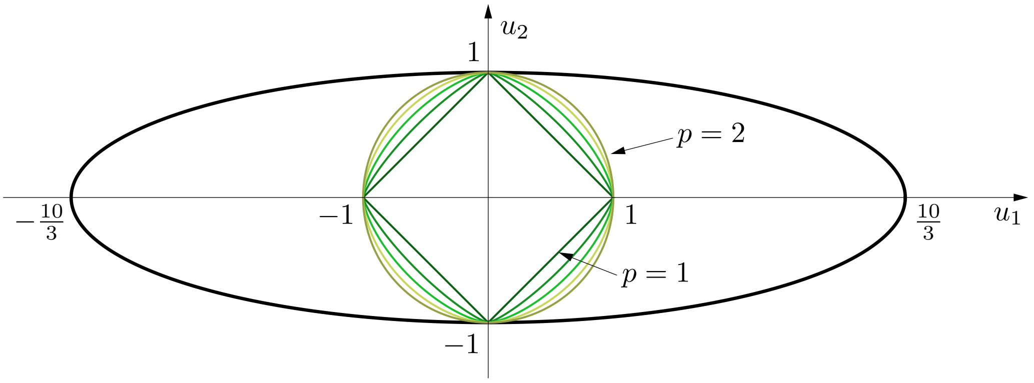


Figure 7.8: Geometry of the optimization problem (7.96). Unit norm balls $|u_1|^p + |u_2|^p = 1$ are shown for $p \in \{1, 1.25, 1.5, 1.75, 2\}$.

Let us look at $\|\boldsymbol{A}^{\ddagger}(\alpha)\|_{\ell^p \to \ell^2} = \max_{\boldsymbol{u} \in \mathbb{R}^2 : \|\boldsymbol{u}\|_p = 1} \|\boldsymbol{A}^{\ddagger}(\alpha)\boldsymbol{u}\|_2$. We can write this as (after squaring the 2-norm)

$$\max_{u_1^p + u_2^p = 1} (\eta^2 + \alpha^2) u_1^2 + u_2^2. \tag{7.96}$$

The optimization problem (7.96) is depicted geometrically in Figure 7.8: consider the ellipse with the equation

$$(\eta^2 + \alpha^2) u_1^2 + u_2^2 = R^2. \tag{7.97}$$

We are searching for the largest R so that there exists a point on this ellipse with the unit ℓ^p-norm. In other words, we are squeezing the ellipse until it touches the ℓ^p-norm ball. The semi-axes of the ellipse are $R/\sqrt{\eta^2 + \alpha^2}$ and R, so that when η and α are both small, the ellipses are elongated along the horizontal axis and the intersection between the squeezed ellipse and the ℓ^p ball will be close to the vertical axis. In fact, for $p \leqslant 2$, one can check[6] that there exists $0 < \beta < 1$ such that, whenever $\eta^2 + \alpha^2 \leqslant \beta^2$, the squeezed ellipse touches the ℓ^p ball only at the points $[0, \pm 1]^\top$ (as seen on Figure 7.8). That is, the maximum is achieved for $R = 1$. The value of the cost function at this maximum is

$$(\eta^2 + \alpha^2) 0^2 + (\pm 1)^2 = 1 \tag{7.98}$$

Therefore, choosing $\eta < \beta$, the value of $\|\boldsymbol{A}^{\ddagger}(\alpha)\|_{\ell^p \to \ell^2}$ is constant for all $\alpha^2 \leqslant \beta^2 - \eta^2$, showing that there are many generalized inverses yielding the same $\ell^p \to \ell^2$ norm as the MPP, which we know achieves the optimum.

---

[6] This is no longer the case for $p > 2$ for the ℓ^p ball is "too smooth" at the point $[0, 1]$: around this point on the ball we have $1 - u_2 = c|u_1|^p(1 + o(u_1)) = o(u_1^2)$.

Conclusion

> It's been a hard day's night,
> and I'd been working like a dog.
> It's been a hard day's night,
> I should be sleeping like a log...
>
> *The Beatles*

While doing the research described in this thesis, ideas for new research were occurring at a steady pace. Writing the thesis was a similar experience: Nearly every page inspired new ideas which seemed more attractive to pursue than the writing itself. Older ideas then kept generating their progeny ideas, while the new ideas generated new ideas of their own, so the number of ideas at page n should not have been too far from the nth Fibonacci number.

Therefore, instead of re-summarizing the content of the thesis, we decided to use this space to outline a research proposal based on our findings and a small selection of the mentioned ideas.

Hearing Room Shapes and Echo Sorting Recently it became clear to us that modeling the room probabilistically could obviate the problems that arise because of the missing echoes, overlapping echoes, noise, and model mismatch. It also seems natural to extend the theory to moving microphones and sources, which quickly brings us into the realm of simultaneous localization and mapping (SLAM) and its vast literature.

Some of this literature looks at time-of-arrival problems similar to the ones that we address by echo sorting. Particularly appealing are the works of Meissner, Leitinger and their colleagues on UWB positioning and SLAM [121, 142]. But the existing works rely on beacons, and the big question is: What can we do with zero fixed infrastructure? That is, what can we do with a standalone device without making assumptions about the space to be mapped except that it contains reflectors? We believe that a fast and practical method should use both the echo sorting and the probabilistic perspective.

Another important line of research is hearing rooms passively. That means reducing the infrastructural demands from Chapter 3 even further by not requiring any active sources. Work in this direction has already begun: Crocco and colleagues [42] have spearheaded the topic and their numerical simulations indicate the feasibility of the passive approach. However, their models are still relatively simple, and the question is how to obtain completely general formulations that efficiently exploit whatever geometric regularities there are?

A different line of work which we did not pursue here are analytic approaches to room hearing. If we think about the problem in terms of the wave or Helmholtz equation on a set of domains, it is not difficult to concoct a cost function involving the room boundary and try to minimize it. Unfortunately, there is no hope for a frontal attack to succeed, as any such cost function has a googol of local minima. Nevertheless, for certain parametric classes of geometries this optimization becomes feasible. The question is then whether we can make it work in practically interesting scenarios. This approach also has potential to take us beyond flat walls required by the image source model.

Crystallography Crystallography has been among the most important disciplines in science for decades now, and it has allowed us to learn the structure of the big molecules that make us up, such as proteins and DNA, and of molecular machines such as ribosomes. In a way, it uncovered the structure of life. Somewhat unexpectedly, the echo sorting toolbox is relevant to determining the molecular structure. Methods such as X-ray or NMR crystallography produce unlabeled sets of distances (perhaps indirectly), and this geometry can be leveraged to develop algorithms [173]. It is often assumed that we know the labeling (it may be obtained by complementary methods and expert knowledge).

But ultimately, the prize is in developing *ab initio* reconstruction algorithms that do the labeling all on their own. A prime example of a crystallography method that could benefit from these algorithms is powder diffraction [222]; the prospect of adapting our results in this context is thrilling.

Expanding the EDM Toolbox EDMs are a powerful (and underused) tool. Many challenges related to EDMs are still to be resolved. We have seen that for small numbers of points (opposite from the current big data trend), it is not easy to get the point set in the correct embedding dimension. Thus one should investigate how to enforce rank in smaller problems, while guaranteeing the optimality.

Another quirk is that to use EDMs, we need actual distances. When we lose synchronization between the sources and the receivers, we can only measure distance differences. We know that this class of problems is also solvable, but as far as we know, there is no nice tool such as an EDM to address distance differences. The geometry changes from circles to hyperbolas, and things become more difficult, so having an EDM-like formulation that is a natural vehicle for algorithm design should unlock numerous new applications.

Information-Theoretic Perspectives Time and time again we work with corrupted (missing, noisy, unlabeled) distances between points. The relevant point sets arise from wave propagation in enclosures. That is, they arise from echoes. An interesting question is how much information can we *fundamentally* extract from these distances knowing that they are noisy, unlabeled and incomplete? There is a limit to how much we can infer, and the way to establish it could be by the tools of information theory. The question here is, *if we had all the computational power we like, what is the most accurate description of the point set that we could obtain according to some reasonable metric, given the corrupted distance data?*

Wideband Diversity In Chapter 4 we have seen that by exploiting the full bandwidth of the sources we can localize multiple sources with only a single microphone. This may initially come as a surprise, but it starts making sense after one reflects upon the problem for a moment. What makes it work? The key is in having a *rich* spectral signature at every candidate location. Drawing parallels with compressed sensing, random spectral signatures seem like good candidates. In a room they are fixed (we cannot design them), but there is a different, bigger issue.

When relying on the room to provide us with this diversity, we make a very strong assumption that the room is known. This is clearly impractical for ad hoc deployments. But what if we could take a room that we know well with ourselves?

Indeed, we can design our spectral signatures if we change the perspective: Imagine putting the microphone (or microphones) on an opportunely designed scattering structure. The idea initially came from thinking about head-related transfer functions. People can identify whether a sound is coming from below or from above the median plane. Knowing that ears are inside the median plane, this ability may be surprising until we give up on omnidirectionality. Our skull, pinna and torso filter sounds coming from down below differently from those coming from up above, and our brains learned how to use these cues to determine the elevation of the sound source.

So the idea is—let us build our own rich spectral signatures by designing *the best of heads* (which may incidentally be random). Then we can place any number of microphones on, or inside this head. One microphone will already suffice for localization, but we can also place *more* than two to make it more robust. Once we localize the sources we also learn the transfer functions which further facilitates source separation. Possibilities are exciting and ample.

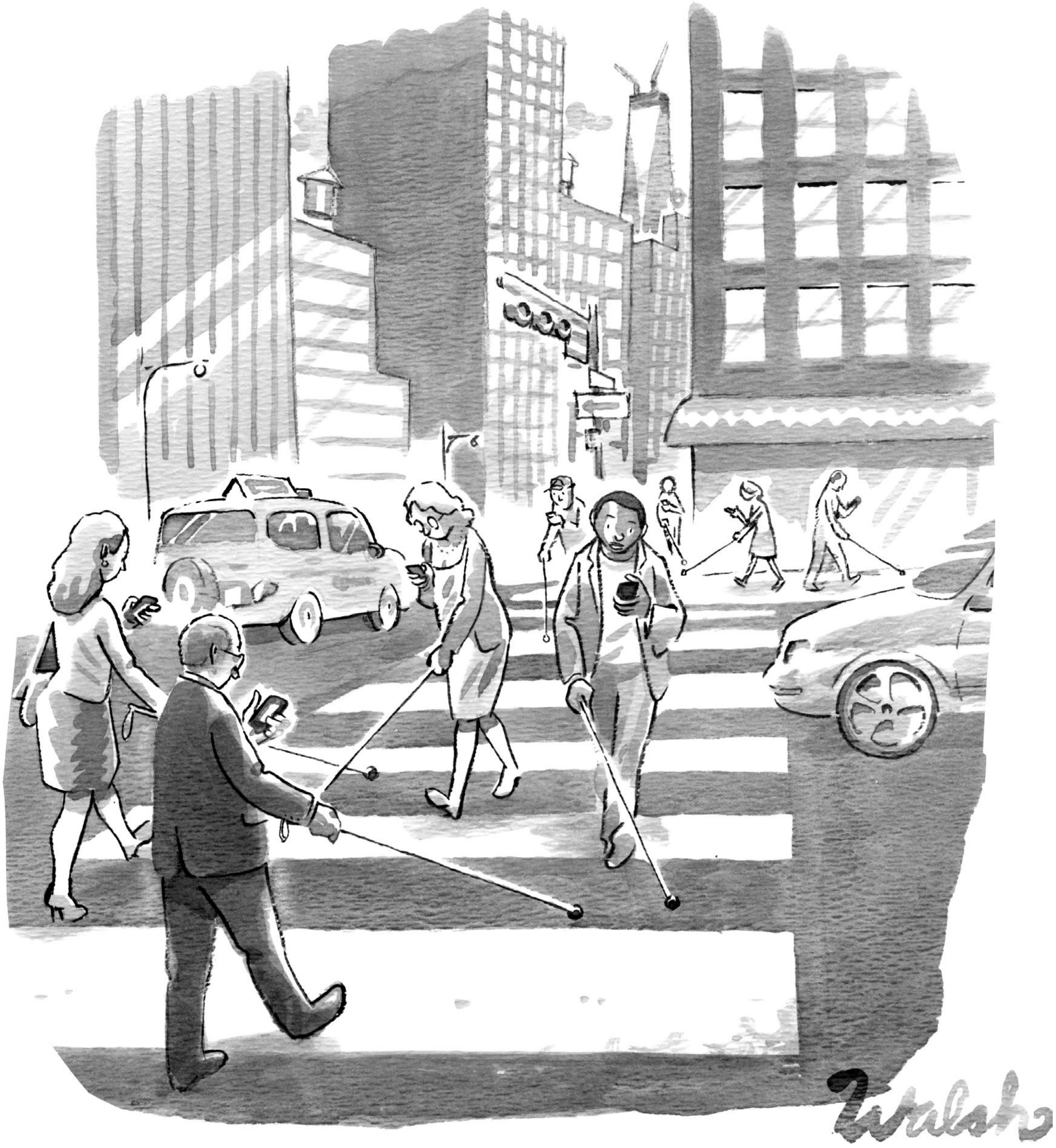
The wideband diversity can also help in sparse sampling in the vein of FRI. We have seen in Chapter 5 how the source localization problem with spherical microphone arrays can be posed as a sparse spherical sampling problem. There we posed it at a single frequency, but exactly the same trick could be used to reduce the number of microphones. There, it will translate into a common nullspace property. This should lead to a complete FRI theory for sampling time-varying signals.

Finally, the Helmholtz equation is one particular (Fourier) transform of the wave equation. Other representations are possible [45], and some may be more suitable for localization and separation problems than sinusoids.

Raking for Source Separation A natural next step in the acoustic raking perspective is to use it to enhance (semi-)blind source separation (in wireless communications there is no severe source separation issue, thanks to the signal design). It has been known that source localization and beamforming (*i.e.*, geometric information) aid in source separation [165]. But what if we are in a reverberant environment, and we know where the early echoes are coming from? We predict that this would significantly improve the separation quality.

Ideally, we could account for the early echoes without knowing where they are coming from *a priori*. Similarly as in room geometry estimation, there must exist methods that would allow to leverage the multipath structure without explicitly estimating it. Maybe by using the image source geometry of early reflections as a form of regularization in solving an inverse problem, with a cleverly designed cost function which we could efficiently optimize.

A Closing Note This brings us to the end of this thesis. It was a long journey through problems and answers about echoes, sound, and beyond, and we hope you do not regret your time investment. As in the beginning of the thesis, we maintain: *It's a good idea to listen to echoes—they're friends!*


Walsh

Bibliography

[1] M. Abramowitz and I. A. Stegun, "Handbook of mathematical functions," National Bureau of Standards, 1972.

[2] P. A. R. Ade, R. W. Aikin, D. Barkats, S. J. Benton, C. A. Bischoff, J. J. Bock, J. A. Brevik, I. Buder, E. Bullock, C. D. Dowell, L. Duband, J. P. Filippini, S. Fliescher, S. R. Golwala, M. Halpern, M. Hasselfield, S. R. Hildebrandt, G. C. Hilton, V. V. Hristov, K. D. Irwin, K. S. Karkare, J. P. Kaufman, B. G. Keating, S. A. Kernasovskiy, J. M. Kovac, C. L. Kuo, E. M. Leitch, M. Lueker, P. Mason, C. B. Netterfield, H. T. Nguyen, R. O'Brient, R. W. Ogburn, A. Orlando, C. Pryke, C. D. Reintsema, S. Richter, R. Schwarz, C. D. Sheehy, Z. K. Staniszewski, R. V. Sudiwala, G. P. Teply, J. E. Tolan, A. D. Turner, A. G. Vieregg, C. L. Wong, K. W. Yoon, and BICEP2 Collaboration, "Detection of B-Mode Polarization at Degree Angular Scales by BICEP2," *Phys. Rev. Lett.*, vol. 112, no. 24, p. 241101, June 2014.

[3] T. Ajdler, L. Sbaiz, and M. Vetterli, "The Plenacoustic Function and Its Sampling," *IEEE Trans. Signal Process.*, vol. 54, no. 10, pp. 3790–3804, 2006.

[4] A. Y. Alfakih, A. Khandani, and H. Wolkowicz, "Solving Euclidean Distance Matrix Completion Problems via Semidefinite Programming," *Computational Optimization and Applications*, vol. 12, no. 1-3, pp. 13–30, 1999.

[5] J. B. Allen and D. A. Berkley, "Image Method For Efficiently Simulating Small-room Acoustics," *J. Acoust. Soc. Am.*, vol. 65, no. 4, pp. 943–950, 1979.

[6] P. Annibale, F. Antonacci, P. Bestagini, A. Brutti, A. Canclini, L. Cristoforetti, J. Filos, E. Habets, W. Kellerman, K. Kowalczyk, A. Lombard, E. Mabande, D. Markovic, P. Naylor, and M. Omologo, "The SCENIC Project: Space-Time Audio Processing for Environment-Aware Acoustic Sensing and Rendering," in Proc. *131st Convention of the Audio Engineering Society.* New York, NY, USA: Audio Engineering Society, 2011.

[7] F. Antonacci, J. Filos, M. R. P. Thomas, E. A. P. Habets, A. Sarti, P. A. Naylor, and S. Tubaro, "Inference of Room Geometry From Acoustic Impulse Responses," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 20, no. 10, pp. 2683–2695, 2012.

[8] F. Antonacci, D. Lonoce, M. Motta, A. Sarti, and S. Tubaro, "Efficient Source Localization and Tracking in Reverberant Environments Using Microphone Arrays," in Proc. *IEEE ICASSP*, Philadelphia, 2005.

[9] P. Audet, "Directional Wavelet Analysis on the Sphere: Application to Gravity and Topography of the Terrestrial Planets," *J. Geophys. Res.*, vol. 116, no. E1, pp. 1–16, 2011.

[10] R. G. Baraniuk, V. Cevher, M. F. Duarte, and C. Hegde, "Model-Based Compressive Sensing," *IEEE Trans. Inf. Theory*, vol. 56, no. 4, pp. 1982–2001, 2010.

[11] R. F. Bass and K. Gröchenig, "Random Sampling of Multivariate Trigonometric Polynomials," *SIAM J. Math. Anal.*, vol. 36, no. 3, pp. 773–795, Jan. 2005.

[12] A. Beck, P. Stoica, and J. Li, "Exact and Approximate Solutions of Source Localization Problems," *IEEE Trans. Signal Process.*, vol. 56, no. 5, pp. 1770–1778, 2008.

[13] A. Ben-Israel and T. N. E. Greville, *Generalized Inverses: Theory and Applications.* New York: Springer, June 2003.

[14] T. Bendory, S. Dekel, and A. Feuer, "Exact Recovery of Dirac Ensembles from the Projection Onto Spaces of Spherical Harmonics," *Constr. Approx.*, pp. 1–25, 2014.

[15] ——, "Super-resolution on the Sphere using Convex Optimization," *arXiv*, Dec. 2014.

[16] J. Benesty, J. Chen, Y. A. Huang, and J. Dmochowski, "On Microphone-Array Beamforming From a MIMO Acoustic Signal Processing Perspective," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 3, pp. 1053–1065, Mar. 2007.

[17] P. Biswas, T. C. Liang, K. C. Toh, Y. Ye, and T. C. Wang, "Semidefinite Programming Approaches for Sensor Network Localization With Noisy Distance Measurements," *IEEE Trans. Autom. Sci. Eng. Trans. Autom. Sci. Eng. Trans. Autom. Sci. Eng. Trans. Autom. Sci. Eng. IEEE Trans. Aut. Sci. Eng.*, vol. 3, no. 4, pp. 360–371, 2006.

[18] I. Borg and P. J. F. Groenen, *Modern Multidimensional Scaling*, Springer Series in Statistics. New York, NY: Springer New York, 2005.

[19] J. Borish, "Extension of the Image Model To Arbitrary Polyhedra," *J. Acoust. Soc. Am.*, vol. 75, no. 6, pp. 1827–1836, 1984.

[20] R. Bott and R. J. Duffin, "On the algebra of networks," 1953.

[21] P. T. Boufounos, P. Smaragdis, and B. Raj, "Joint Sparsity Models for Wideband Array Processing," *Wavelets and Sparsity XIV*, vol. 8138, pp. 81 380K–81 380K–10, Sept. 2011.

[22] A. Bourrier, M. E. Davies, T. Peleg, P. Perez, and R. Gribonval, "Fundamental Performance Limits for Ideal Decoders in High-Dimensional Linear Inverse Problems," *IEEE Trans. Inf. Theory*, pp. 1–1, 2014.

[23] M. Boutin and G. Kemper, "On Reconstructing N-Point Configurations From the Distribution of Distances or Areas," *Adv. Appl. Math.*, vol. 32, no. 4, pp. 709–735, May 2004.

[24] J. S. Bradley, R. D. Reich, and S. G. Norcross, "On the Combined Effects of Signal-to-Noise Ratio and Room Acoustics on Speech Intelligibility," *J. Acoust. Soc. Am.*, 1999.

[25] J. S. Bradley, H. Sato, and M. Picard, "On the Importance of Early Reflections for Speech in Rooms," *J. Acoust. Soc. Am.*, vol. 113, no. 6, p. 3233, 2003.

[26] M. W. Browne, "The Young-Householder Algorithm and the Least Squares Multidimensional Scaling of Squared Distances," *J. Classification*, vol. 4, no. 2, pp. 175–190, 1987.

[27] H. P. Bucker, “Use of Calculated Sound Fields and Matched-Field Detection to Locate Sound Sources in Shallow Water,” *J. Acoust. Soc. Am.*, vol. 59, no. 2, pp. 368–373, 1976.

[28] T. Bulow, “Spherical Diffusion for 3D Surface Smoothing,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 12, pp. 1650–1654, 2004.

[29] J. A. Cadzow, “Signal Enhancement—A Composite Property Mapping Algorithm,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 1, pp. 49–62, 1988.

[30] E. J. Candès and C. Fernandez-Granda, “Towards a Mathematical Theory of Super-resolution,” *Commun. Pur. Appl. Math.*, vol. 67, no. 6, pp. 906–956, June 2014.

[31] J. Capon, “High-Resolution Frequency-Wavenumber Spectrum Analysis,” *Proc. IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.

[32] B. D. Carlson, “Covariance Matrix Estimation Errors and Diagonal Loading in Adaptive Arrays,” *IEEE Trans. Aerosp. Electron. Syst.*, vol. 24, no. 4, pp. 397–401, July 1988.

[33] P. G. Casazza, A. Heinecke, F. Krahmer, and G. Kutyniok, “Optimally Sparse Frames,” *IEEE Trans. Inf. Theory*, vol. 57, no. 11, pp. 7279–7287.

[34] S. J. Chapman, “Drums that sound the same,” *Am. Math. Mon.*, vol. 102, no. 2, pp. 124–138, Feb. 1995.

[35] G. Chardon and L. Daudet, “Narrowband Source Localization in an Unknown Reverberant Environment Using Wavefield Sparse Decomposition,” in Proc. *IEEE ICASSP*, pp. 9–12, 2012.

[36] G. Chardon, “Approximations parcimonieuses et problèmes inverses en acoustique ,” Ph.D. dissertation, Université Pierre et Marie Curie, July 2012.

[37] K. W. Cheung, W. K. Ma, and H. C. So, “Accurate Approximation Algorithm for TOA-Based Maximum Likelihood Mobile Location Using Semidefinite Programming,” in Proc. *IEEE ICASSP*, pp. ii–145–8 vol.2. IEEE, 2004.

[38] O. Christensen, *Frames and Bases*, An Introductory Course. Boston: Springer Science & Business Media, June 2008.

[39] A. Comte, “Cours de Philosophie Positive: La Philosophie Astronomique et la Philosophie de la Physique,” 1835.

[40] ——, *The Positive Philosophy.* Kitchener: Batoche Books, 2000.

[41] M. Crocco, A. Del Bue, M. Bustreo, and V. Murino, “A Closed Form Solution to the Microphone Position Self-Calibration Problem,” in Proc. *IEEE ICASSP*, pp. 2597–2600, 2012.

[42] M. Crocco, A. Trucco, V. Murino, and A. Del Bue, “Towards Fully Uncalibrated Room Reconstruction with Sound,” in Proc. *EUSIPCO*, pp. 910–914. Lisabon: IEEE, 2014.

[43] M. Crocco, A. D. Bue, and V. Murino, “A Bilinear Approach to the Position Self-Calibration of Multiple Sensors,” *IEEE Trans. Signal Process.*, vol. 60, no. 2, pp. 660–673, 2012.

[44] J. Dattorro, *Convex Optimization & Euclidean Distance Geometry.* Meboo, 2011.

[45] L. Demanet, “Curvelets, Wave Atoms, and Wave Equations,” Ph.D. dissertation, California Institute of Technology, 2006.

[46] S. Deslauriers-Gauthier and P. Marziliano, “Spherical Finite Rate of Innovation Theory for the Recovery of Fiber Orientations,” in Proc. *IEEE EMBC*, pp. 2294–2297, San Diego, CA, USA, 2012.

[47] ——, “Sampling Signals With a Finite Rate of Innovation on the Sphere,” *IEEE Trans. Signal Process.*, vol. 61, no. 18, pp. 4552–4561, 2013.

[48] ——, “Sampling Great Circles at Their Rate of Innovation,” in Proc. *SPIE Wavelets and Sparsity*, D. Van De Ville, V. K. Goyal, and M. Papadakis, eds. San Diego, CA, USA: International Society for Optics and Photonics, Sept. 2013.

[49] P. E. Dewdney, P. J. Hall, R. T. Schilizzi, and T. J. L. W. Lazio, “The Square Kilometre Array,” *Proc. IEEE*, vol. 97, no. 8, pp. 1482–1496, June 2009.

[50] J. H. DiBiase, H. F. Silverman, and M. S. Brandstein, “Robust Localization in Reverberant Rooms,” in *Microphone Arrays.* Berlin, Heidelberg: Springer Berlin Heidelberg, 2001, pp. 157–180.

[51] I. Dokmanić, L. Daudet, and M. Vetterli, “How to Localize Ten Microphones in One Fingersnap,” in Proc. *EUSIPCO*, Lisbon, 2014.

[52] I. Dokmanić, M. Kolundžija, and M. Vetterli, “Beyond Moore-Penrose: Sparse pseudoinverse,” in Proc. *IEEE ICASSP*, pp. 6526–6530, 2013.

[53] I. Dokmanić and Y. M. Lu, “Sampling Sparse Signals on the Sphere: Algorithms and Applications,” *arXiv*, Feb. 2015.

[54] I. Dokmanić, Y. M. Lu, and M. Vetterli, “Can One Hear the Shape of a Room: The 2-D Polygonal Case,” in Proc. *IEEE ICASSP*, pp. 321–324, Prague, 2011.

[55] I. Dokmanić, R. Parhizkar, J. Ranieri, and M. Vetterli, “Euclidean Distance Matrices: Essential Theory, Algorithms, and Applications,” *IEEE Signal Process. Mag.*, 2015.

[56] I. Dokmanić, R. Parhizkar, A. Walther, Y. M. Lu, and M. Vetterli, “Acoustic Echoes Reveal Room Shape,” *Proc. Natl. Acad. Sci.*, vol. 110, no. 30, June 2013.

[57] I. Dokmanić and D. Petrinović, “Convolution on the n-Sphere With Application to PDF Modeling,” *IEEE Trans. Signal Process.*, vol. 58, no. 3, pp. 1157–1170, Mar. 2010.

[58] I. Dokmanić, J. Ranieri, A. Chebira, and M. Vetterli, “Sensor Networks for Diffusion Fields: Detection of Sources in Space and Time,” in Proc. *Allerton*, pp. 1552–1558, 2011.

[59] I. Dokmanić, J. Ranieri, and M. Vetterli, “Relax and Unfold: Microphone Localization with Euclidean Distance Matrices,” in Proc. *EUSIPCO*, Nice, 2015.

[60] I. Dokmanić, R. Scheibler, and M. Vetterli, “Raking the Cocktail Party,” *IEEE J. Sel. Top. Signal Process.*, vol. PP, no. 99, pp. 1–12, 2015.

[61] I. Dokmanić and M. Vetterli, "Room Helps: Acoustic Localization with Finite Elements," in Proc. *IEEE ICASSP*, pp. 2617–2620. IEEE, 2012.

[62] D. L. Donoho, "Compressed Sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[63] P. L. Dragotti and F. Homann, "Sampling Signals with Finite Rate of Innovation in the Presence of Noise," in Proc. *IEEE ICASSP*, pp. 2941–2944. Taipei, Taiwan: IEEE, 2009.

[64] C. Drane, M. Macnaughtan, and C. Scott, "Positioning GSM Telephones," *IEEE Communications Magazine*, vol. 36, no. 4, pp. 46–54, 59, Apr. 1998.

[65] J. R. Driscoll and D. M. Healy, "Computing Fourier Transforms and Convolutions on the 2-Sphere," *Adv. Appl. Math.*, vol. 15, no. 2, pp. 202–250, June 1994.

[66] T. A. Driscoll, "Eigenmodes of Isospectral Drums," *SIAM Rev.*, vol. 39, no. 1, pp. 1–17, 1997.

[67] D. Duffy, *Green's Functions with Applications.* Chapman and Hall/CRC, 2001.

[68] J. Eckstein and D. P. Bertsekas, "On the Douglas–Rachford Splitting Method and the Proximal Point Algorithm for Maximal Monotone Operators," *Math. Program.*, vol. 55, no. 1-3, pp. 293–318, 1992.

[69] A. Edelman, "Eigenvalues and Condition Numbers of Random Matrices," *SIAM J. Matrix Anal. Appl.*, vol. 9, no. 4, pp. 543–560, Oct. 1988.

[70] Y. C. Eldar, P. Kuppinger, and H. Bolcskei, "Block-Sparse Signals: Uncertainty Relations and Efficient Recovery," *IEEE Trans. Signal Process.*, vol. 58, no. 6, pp. 3042–3054, May 2010.

[71] K. F. Evans, "The Spherical Harmonics Discrete Ordinate Method for Three-Dimensional Atmospheric Radiative Transfer," *J. Atmos. Sci.*, vol. 55, no. 3, pp. 429–446, 1998.

[72] A. Farina, "Simultaneous Measurement of Impulse Response and Distortion with a Swept-Sine Technique," in Proc. *Audio Engineering Society Convention 108*, pp. 1–24, 2000.

[73] S. J. Farlow, *Partial Differential Equations for Scientists and Engineers.* Courier Corporation, Mar. 2012.

[74] O. L. I. Frost, "An Algorithm for Linearly Constrained Adaptive Array Processing," *Proc. IEEE*, vol. 60, no. 8, pp. 926–935, 1972.

[75] K. Furuya, "Noise Reduction and Dereverberation using Correlation Matrix Based on the Multiple-Input/Output Inverse-Filtering Theorem (MINT)," in Proc. *Intl. Workshop on HSC*, pp. 59–62, Kyoto, Japan, 2001.

[76] N. Gaffke and R. Mathar, "A Cyclic Projection Algorithm via Duality," *Metrika*, vol. 36, no. 1, pp. 29–54, 1989.

[77] N. D. Gaubitch, W. B. Kleijn, and R. Heusdens, "Auto-Localization in Ad-Hoc Microphone Arrays," in Proc. *IEEE ICASSP*, pp. 106–110. Vancouver: IEEE, 2013.

[78] W. Glunt, T. Hayden, and W. Liu, "The Embedding Problem for Predistance Matrices," *Bull. Math. Biol.*, vol. 53, no. 5, pp. 769–796, 1991.

[79] M. M. Goodwin and G. W. Elko, "Constant Beamwidth Beamforming," in Proc. *IEEE ICASSP*, pp. 169–172 vol.1. IEEE, 1993.

[80] C. Gordon and D. Webb, "You Can't Hear the Shape of a Drum," *Am. Sci.*, vol. 84, pp. 46–55, 1996.

[81] C. Gordon, D. Webb, and S. Wolpert, "Isospectral Plane Domains and Surfaces via Riemannian Orbifolds," *Invent. Math.*, vol. 110, no. 1, pp. 1–22, Dec. 1992.

[82] C. Gordon, D. L. Webb, and S. Wolpert, "One Cannot Hear the Shape of a Drum," *Bull. Amer. Math. Soc*, vol. 27, no. 1, pp. 134–138, 1992.

[83] J. C. Gower, "Euclidean Distance Geometry," *Math. Sci.*, vol. 7, no. 1, pp. 1–14, 1982.

[84] M. Grant and S. Boyd, "Graph Implementations for Nonsmooth Convex Programs," in *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, V. Blondel, S. Boyd, and H. Kimura, eds. Springer-Verlag Limited, 2008, pp. 95–110, http://stanford.edu/~boyd/graph_dcp.html.

[85] ——, "CVX: Matlab Software for Disciplined Convex Programming, version 2.1," http://cvxr.com/cvx, Mar. 2014.

[86] R. Gribonval and M. Nielsen, "Highly Sparse Representations From Dictionaries Are Unique and Independent of the Sparseness Measure," *Appl. Comput. Harmon. Anal.*, vol. 22, no. 3, pp. 335–355, May 2007.

[87] S. R. Gujarathi, C. L. Farrow, C. Glosser, L. Granlund, and P. M. Duxbury, "Ab-Initio Reconstruction of Complex Euclidean Networks in Two Dimensions," *Physical Review E*, vol. 89, no. 5, 2014.

[88] E. Habets, J. Benesty, I. Cohen, S. Gannot, and J. Dmochowski, "New Insights Into the MVDR Beamformer in Room Acoustics," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 1, pp. 158–170, Jan. 2010.

[89] J. Hadamard, "Sur Les Problémes Aux Dérivées Partielles Et Leur Signification Physique," *Princeton University Bulletin*, 1902.

[90] M. Harris, "The Way Through the Flames," *IEEE Spectrum*, vol. 50, no. 9, pp. 30–35, 2013.

[91] S. Haykin and Z. Chen, "The Cocktail Party Problem," *Neural Comput.*, vol. 17, no. 9, pp. 1875–1902, 2005.

[92] W. Herbordt and W. Kellermann, "Adaptive Beamforming for Audio Signal Acquisition," in *Adaptive Signal Processing.* Berlin, Heidelberg: Springer, Feb. 2003, pp. 155–194.

[93] R. A. Horn and C. R. Johnson, *Matrix Analysis.* Cambridge University Press, Oct. 2012.

[94] F. Ihlenburg, *Finite Element Analysis of Acoustic Scattering.* Springer Science & Business Media, Aug. 1998.

[95] ITU-T P.862 Amendment 2, "Reference Implementations and Conformance Testing for ITU-T Recs P.862, P.862.1 and P.862.2," 11 2005. [Online]. Available: https://www.itu.int/rec/T-REC-P.862-200511-I!Amd2/en.

[96] E.-E. Jan, P. Svaizer, and J. L. Flanagan, "Matched-Filter Processing of Microphone Array for Spatial Volume Selectivity," *Proc. IEEE ISCAS*, vol. 2, pp. 1460–1463, 1995.

[97] N. Jarosik, C. L. Bennett, J. Dunkley, B. Gold, M. R. Greason, M. Halpern, R. S. Hill, G. Hinshaw, A. Kogut, E. Komatsu, D. Larson, M. Limon, S. S. Meyer, M. R. Nolta, N. Odegard, L. Page, K. M. Smith, D. N. Spergel, G. S. Tucker, J. L. Weiland, E. Wollack, and E. L. Wright, "Seven-year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Sky Maps, Systematic Errors, and Basic Results," *Astrophys. J. Suppl. Ser.*, vol. 192, no. 2, pp. 1–15, Feb. 2011.

[98] D. P. Jarrett, E. A. P. Habets, M. R. P. Thomas, and P. A. Naylor, "Rigid Sphere Room Impulse Response Simulation: Algorithm and Applications," *J. Acoust. Soc. Am.*, vol. 132, no. 3, pp. 1462–1472, 2012.

[99] M. Kac, "Can One Hear the Shape of a Drum," *Am. Math. Mon.*, vol. 73, pp. 1–23, 1966.

[100] S. Kaczmarz, "Angenäherte Auflösung von Systemen linearer Gleichungen," *Bulletin International de l'Académie Polonaise des Sciences et des Lettres*, vol. 35, pp. 355–357, 1937.

[101] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation theory*, 1998.

[102] J. B. Keller, "The Scope of the Image Method," *Commun. Pur. Appl. Anal.*, vol. 6, no. 4, pp. 505–512, 1953.

[103] ——, "Geometrical Theory of Diffraction," *J. Opt. Soc. Am. B.*, vol. 52, pp. 116–130, Jan. 1962.

[104] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix Completion From a Few Entries," *IEEE Trans. Inf. Theory*, vol. 56, no. 6, pp. 2980–2998, June 2010.

[105] ——, "Matrix Completion from Noisy Entries," *arXiv*, Apr. 2012.

[106] B. H. Khalaj, A. Paulraj, and T. Kailath, "2D RAKE Receivers for CDMA Cellular Systems," in Proc. *IEEE GLOBECOM*, pp. 400–404. IEEE, 1994.

[107] Z. Khalid, R. A. Kennedy, and J. D. McEwen, "An Optimal-Dimensionality Sampling Scheme on the Sphere With Fast Spherical Harmonic Transforms," *IEEE Trans. Signal Process.*, vol. 62, no. 17, pp. 4597–4610, 2014.

[108] S. Kitić, N. Bertin, and R. Gribonval, "Hearing Behind Walls: Localizing Sources in the Room Next Door with Cosparsity," *IEEE ICASSP*, pp. 3087–3091, 2014.

[109] S. Kitić, L. Albera, N. Bertin, and R. Gribonval, "Physics-Driven Inverse Problems Made Tractable with Cosparse Regularization," Mar. 2015. [Online]. Available: https://hal.inria.fr/hal-01133087.

[110] J. Kovačević and A. Chebira, *An Introduction to Frames.* Now Publishers Inc, 2008.

[111] F. Krahmer, G. Kutyniok, and J. Lemvig, “Sparsity and Spectral Properties of Dual Frames,” *Linear Algebra Appl.*, pp. 1–17, Dec. 2012.

[112] H. Krim and M. Viberg, “Two Decades of Array Signal Processing Research: the Parametric Approach,” *IEEE Signal Process. Mag.*, vol. 13, no. 4, pp. 67–94, July 1996.

[113] N. Krislock and H. Wolkowicz, “Euclidean Distance Matrices and Applications,” in *Handbook on Semidefinite, Conic and Polynomial Optimization.* Boston, MA: Springer US, Jan. 2012, pp. 879–914.

[114] J. B. Kruskal, “Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis,” *Psychometrika*, vol. 29, no. 1, pp. 1–27, Mar. 1964.

[115] Y. Kuang, S. Burgess, A. Torstensson, and K. Astrom, “A Complete Characterization and Solution to the Microphone Position Self-Calibration Problem,” in Proc. *IEEE ICASSP*, pp. 3875–3879. IEEE, 2013.

[116] P. Kurmann, *La Cathédrale Notre-Dame de Lausanne: Monument Européen, Temple Vaudois* . Editions La Bibliothèque des Arts, 2012.

[117] H. Kuttruff, *Room Acoustics*, 5th ed. CRC Press, June 2009.

[118] M. Lammers, A. M. Powell, and Ö. Yılmaz, “Alternative Dual Frames for Digital-to-Analog Conversion in Sigma–Delta Quantization,” *Adv Comput Math*, vol. 32, no. 1, pp. 73–102, July 2008.

[119] E. G. Larsson and D. Danev, “Accuracy Comparison of LS and Squared-Range LS for Source Localization,” *IEEE Trans. Signal Process.*, vol. 58, no. 2, pp. 916–923, 2010.

[120] J. Le Roux, P. T. Boufounos, K. Kang, and J. R. Hershey, “Source Localization in Reverberant Environments Using Sparse Optimization,” in Proc. *IEEE ICASSP*, pp. 4310–4314. IEEE, 2013.

[121] E. Leitinger, P. Meissner, M. Lafer, and K. Witrisal, “Simultaneous Localization and Mapping using Multipath Channel Information,” in Proc. *IEEE ICC*, June 2015.

[122] J. Leng, D. Han, and T. Huang, “Optimal Dual Frames for Communication Coding With Probabilistic Erasures,” *IEEE Trans. Signal Process.*, vol. 59, no. 11, pp. 5380–5389.

[123] A. D. Lewis, “A Top Nine List: Most Popular Induced Matrix Norms,” Queen’s University, Kingston, Ontario, Tech. Rep., 2010.

[124] H. Liu, H. Darabi, P. Banerjee, and J. Liu, “Survey of Wireless Indoor Positioning Techniques and Systems,” *IEEE Trans. Systems, Man, and Cybernetics*, vol. 37, no. 6, pp. 1067–1080, Nov. 2007.

[125] J. Liu and J. Ye, “Efficient ℓ^1/ℓ^q Norm Regularization,” Sept. 2010.

[126] Y. Liu, T. Mi, and S. Li, “Compressed Sensing With General Frames via Optimal-Dual-Based -Analysis,” *IEEE Trans. Inf. Theory*, vol. 58, no. 7, pp. 4201–4214, 2012.

[127] J. Lochner and J. F. Burger, “The Influence of Reflections on Auditorium Acoustics,” *J. Sound Vib.*, vol. 1, no. 4, pp. 426–454, 1964.

[128] T. Lokki and V. Pulkki, “Evaluation of Geometry-based Parametric Auralization,” *AES22: International Conference: Virtual, Synthetic, and Entertainment Audio*, 2002.

[129] J. Lopez and D. Han, “Optimal Dual Frames for Erasures ,” *Linear Algebra Appl.*, vol. 432, no. 1, pp. 471–482, Jan. 2010.

[130] Y. M. Lu, P. L. Dragotti, and M. Vetterli, “Localizing Point Sources in Diffusion Fields From Spatiotemporal Samples,” in Proc. *SampTA*, Signapore, 2011.

[131] Y. M. Lu and M. Vetterli, “Distributed Spatio-Temporal Sampling of Diffusion Fields From Sparse Instantaneous Sources,” in Proc. *IEEE CAMSAP*, Aruba, 2009.

[132] R. H. Macphie and E. H. Okongwu, “Spherical Harmonics and Earth-Rotation Synthesis in Radio Astronomy,” *IEEE Trans. Antennas Propag.*, vol. 23, no. 3, pp. 386–391, 1975.

[133] D. Malioutov, M. Cetin, and A. S. Willsky, “A Sparse Signal Reconstruction Perspective for Source Localization with Sensor Arrays,” *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 3010–3022, Aug. 2005.

[134] D. M. Malioutov, “A Sparse Signal Reconstruction Perspective for Source Localization with Sensor Arrays (Master Thesis),” 2003.

[135] D. E. Manolakis, “Efficient Solution and Performance Analysis of 3-D Position Estimation by Trilateration,” *IEEE Trans. Aerosp. Electron. Syst.*, vol. 32, no. 4, pp. 1239–1248, 1996.

[136] I. Maravic and M. Vetterli, “Exact Sampling Results for Some Classes of Parametric Non-bandlimited 2-D Signals,” *IEEE Trans. Signal Process.*, vol. 52, no. 1, pp. 175–189, Jan. 2004.

[137] ——, “Sampling and Reconstruction of Signals With Finite Rate of Innovation in the Presence of Noise ,” *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 2788–2805, Aug. 2005.

[138] D. Marković, F. Antonacci, A. Sarti, and S. Tubaro, “Estimation of Room Dimensions From a Single Impulse Response,” *WASPAA*, pp. 1–4, 2013.

[139] P. Marziliano, M. Vetterli, and T. Blu, “Sampling and Exact Reconstruction of Bandlimited Signals with Additive Shot Noise,” *IEEE Trans. Inf. Theory*, vol. 52, no. 5, pp. 2230–2233, May 2006.

[140] W. McCrea and F. Whipple, “Random Paths in Two and Three Dimensions,” in Proc. *Proc. Roy. Soc. Edinburgh*, vol. 60, pp. 281–298, 1940.

[141] J. D. McEwen and Y. Wiaux, “A Novel Sampling Theorem on the Sphere,” *IEEE Trans. Signal Process.*, vol. 59, no. 12, pp. 5876–5887, 2011.

[142] P. Meissner, “Multipath-Assisted Indoor Positioning,” Ph.D. dissertation, TU Graz, 2014.

[143] P. Meissner and K. Witrisal, “Multipath-Assisted Single-Anchor Indoor Localization in an Office Environment,” in Proc. *IWSSIP*, pp. 22–25, 2012.

[144] P. Meissner, C. Steiner, and K. Witrisal, “UWB Positioning with Virtual Anchors and Floor Plan Information,” in Proc. *WPNC*, pp. 150–156. IEEE, 2010.

[145] J. Meyer and G. Elko, “A Highly Scalable Spherical Microphone Array Based on an Orthonormal Decomposition of the Soundfield,” in Proc. *IEEE ICASSP*, pp. 1781–1784. Orlando, FL, USA: IEEE, 2002.

[146] N. Minamide and K. Nakamura, “A Restricted Pseudoinverse and Its Application to Constrained Minima,” *SIAM J. Appl. Math.*, 1970.

[147] L. Mirsky, “Symmetric Gauge Functions and Unitarily Invariant Norms,” *Q. J. Math. Q. J. Math. Q. J. Math. Q. J. Math.*, vol. 11, no. 1, pp. 50–59, 1960.

[148] M. Miyoshi and Y. Kaneda, “Inverse Filtering of Room Acoustics,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 36, no. 2, pp. 145–152, 1988.

[149] D. Model and M. Zibulevsky, “Signal reconstruction in sensor arrays using sparse representations,” *Signal Process.*, vol. 86, no. 3, pp. 624–638, Mar. 2006.

[150] A. Moiola, R. Hiptmair, and I. Perugia, “Plane Wave Approximation of Homogeneous Helmholtz Solutions,” *Z. Angew. Math. Phys.*, vol. 62, no. 5, pp. 809–837, 2011.

[151] ——, “Vekua theory for the Helmholtz operator,” *Z. Angew. Math. Phys.*, vol. 62, no. 5, pp. 779–807, 2011.

[152] A. H. Moore, M. Brookes, and P. A. Naylor, “Room Geometry Estimation From a Single Channel Acoustic Impulse Response,” in Proc. *EUSIPCO*, pp. 1–5, 2013.

[153] E. H. Moore, “On the Reciprocal of the General Algebraic Matrix,” *Bulletin of the American Mathematical Society*, vol. 26, pp. 394–395, 1920.

[154] P. M. Morse and U. K. Ingard, *Theoretical Acoustics.* New Jersey: Princeton University Press, 1968.

[155] A. F. Naguib, “Space-Time Receivers for CDMA Multipath Signals,” in Proc. *Proc. IEEE ICC*, pp. 304–308. Montreal: IEEE, 1997.

[156] D. Needell and J. A. Tropp, “CoSaMP: Iterative Signal Recovery From Incomplete and Inaccurate Samples,” *Appl. Comput. Harmon. Anal.*, vol. 26, no. 3, pp. 301–321, May 2009.

[157] T. G. Newman and P. L. Odell, “On the Concept of a $p-q$ Generalized Inverse of a Matrix,” *SIAM J. Appl. Math.*, vol. 17, no. 3, pp. 520–525, 1969.

[158] T. Nowakowski, L. Daudet, and J. de Rosny, “Microphone Array Position Calibration in the Frequency Domain Using a Single Unknown Source ,” in Proc. *IEEE ICASSP*, Brisbane, 2015.

[159] O. Öçal, I. Dokmanić, and M. Vetterli, “Source Localization and Tracking in Non-Convex Rooms,” *Proc. IEEE ICASSP*, 2014.

[160] A. E. O’Donovan, R. Duraiswami, and D. N. Zotkin, “Automatic Matched Filter Recovery via the Audio Camera,” in Proc. *Proc. IEEE ICASSP*, pp. 2826–2829. Dallas: IEEE, 2010.

[161] T. E. Oliphant, “Python for Scientific Computing,” *IEEE Comput. Sci. Eng.*, vol. 9, no. 3, pp. 10–20, 2007.

[162] R. Parhizkar, “Euclidean Distance Matrices: Properties, Algorithms and Applications,” Ph.D. dissertation, Ecole Polytechnique Federale de Lausanne (EPFL), Lausanne, 2013.

[163] R. Parhizkar, I. Dokmanić, and M. Vetterli, “Single-Channel Indoor Microphone Localization,” in Proc. *IEEE ICASSP*, pp. 1434–1438, 2014.

[164] N. Parikh and S. Boyd, “Proximal Algorithms,” *Foundations and Trends in Optimization*, vol. 1, no. 3, pp. 123–231, 2014.

[165] L. C. Parra and C. V. Alvino, “Geometric Source Separation: Merging Convolutive Source Separation with Geometric Beamforming,” *IEEE Trans. Speech Audio Process.*, vol. 10, no. 6, pp. 352–362, Sept. 2002.

[166] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, “Orthogonal Matching Pursuit: Recursive Function Approximation with Applications to Wavelet Decomposition,” in Proc. *Asilomar Conf. Signals, Syst., Comput.*, pp. 40–44. IEEE Comput. Soc. Press, Nov. 1993.

[167] R. Penrose, “A Generalized Inverse for Matrices,” *Math. Proc. Camb. Phil. Soc.*, vol. 51, no. 03, p. 406, Oct. 2008.

[168] N. Perraudin, N. Holighaus, P. L. Søndergaard, and P. Balazs, “Designing Gabor Windows Using Convex Optimization,” *arXiv*, Jan. 2014.

[169] R. L. Pickholtz, D. L. Schilling, and L. B. Milstein, “Theory of Spread-Spectrum Communications–A Tutorial,” *IEEE Trans. Commun.*, vol. 30, no. 5, pp. 855–884, 1982.

[170] M. Pollefeys and D. Nister, “Direct Computation of Sound and Microphone Locations From Time-Difference-of-Arrival Data,” in Proc. *IEEE ICASSP*, pp. 2445–2448. Las Vegas: IEEE, 2008.

[171] R. Price and P. E. Green, “A Communication Technique for Multipath Channels,” in Proc. *Proceedings of the IRE*, pp. 555–570. IEEE, 1958.

[172] A. Quattoni, X. Carreras, M. Collins, and T. Darrell, “An Efficient Projection for $\ell_{1,\infty}$ Regularization,” in Proc. *ICML ’09*, pp. 857–864. New York, New York, USA: ACM, June 2009.

[173] J. Ranieri, “Sensing the Real World,” Ph.D. dissertation, Ecole Polytechnique Federale de Lausanne (EPFL), 2014.

[174] V. C. Raykar and R. Duraiswami, “Automatic Position Calibration of Multiple Microphones,” in Proc. *IEEE ICASSP*, pp. 69–72, Montreal, 2004.

[175] F. Ribeiro, D. E. Ba, and C. Zhang, “Turning Enemies Into Friends: Using Reflections to Improve Sound Source Localization,” *Proceedings of the 2010 IEEE International Conference on Multimedia and Expo, ICME 2010, 19-23 July 2010*, 2010.

[176] F. Ribeiro, D. A. Florencio, D. E. Ba, and C. Zhang, “Geometrically Constrained Room Modeling With Compact Microphone Arrays,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 20, no. 5, pp. 1449–1460, 2012.

[177] F. Ribeiro, C. Zhang, D. A. Florencio, and D. E. Ba, "Using Reverberation to Improve Range and Elevation Discrimination for Small Array Sound Source Localization," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 18, no. 7, pp. 1781–1792, 2010.

[178] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual Evaluation of Speech Quality (PESQ)—A New Method for Speech Quality Assessment of Telephone Networks and Codecs," in Proc. *IEEE ICASSP*, pp. 749–752, Salt Lake City, UT, 2001.

[179] L. D. Rosenblum, M. S. Gordon, and L. Jarquin, "Echolocating Distance by Moving and Stationary Listeners," *Ecol. Psychol.*, vol. 12, no. 3, pp. 181–206, July 2000.

[180] M. Rudelson and R. Vershynin, "The Littlewood–Offord Problem and Invertibility of Random Matrices," *Adv. Math.*, vol. 218, no. 2, pp. 600–633, June 2008.

[181] R. T. S, *Wireless Communications: Principles And Practice.* Pearson Education India, Sept. 2010.

[182] J. M. Sachar, H. F. Silverman, and W. R. Patterson, "Microphone Position and Gain Calibration for a Large-Aperture Microphone Array," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 1, pp. 42–52, Jan. 2005.

[183] R. Scheibler, I. Dokmanić, R, and M. Vetterli, "Raking Echoes in the Time Domain," *Accepted to IEEE ICASSP*, 2015.

[184] R. Scheibler, I. Dokmanić, and M. Vetterli, "Raking Echoes in the Time Domain," in Proc. *IEEE ICASSP*, Brisbane, 2014.

[185] P. H. Schönemann, "A Solution of the Orthogonal Procrustes Problem With Applications to Orthogonal and Oblique Rotation," Ph.D. dissertation, University of Illinois at Urbana-Champaign, 1964.

[186] ——, "On Metric Multidimensional Unfolding," *Psychometrika*, vol. 35, no. 3, pp. 349–366, 1970.

[187] P. Shukla and P. L. Dragotti, "Sampling Schemes for Multidimensional Signals With Finite Rate of Innovation," *IEEE Trans. Signal Process.*, vol. 55, no. 7, pp. 3670–3686, 2007.

[188] J. J. Shynk, "Frequency-Domain and Multirate Adaptive Filtering," *IEEE Signal Process. Mag.*, 1992.

[189] F. J. Simons, F. A. Dahlen, and M. A. Wieczorek, "Spatiospectral Concentration on a Sphere," *SIAM Rev.*, vol. 48, no. 3, pp. 504–536, 2006.

[190] J. O. Smith III and J. S. Abel, "Closed-Form Least-Squares Source Location Estimation From Range-Difference Measurements," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 35, no. 12, 1987.

[191] A. M.-C. So and Y. Ye, "Theory of Semidefinite Programming for Sensor Network Localization," *Math. Program.*, vol. 109, no. 2-3, pp. 367–384, Mar. 2007.

[192] S. Sra, "Fast Projections Onto Mixed-Norm Balls with Applications," *Data Min. Knowl. Disc.*, vol. 25, no. 2, pp. 358–377, Sept. 2012.

[193] G. Strang, *Computational Science and Engineering.* Cambridge University Press, Nov. 2007.

[194] G. Strang and G. J. Fix, *An Analysis of the Finite Element Method*, Prentice-Hall series in automatic computation. New Jersey: Prentice-Hall, 1973.

[195] T. Strohmer and R. Vershynin, "A Randomized Kaczmarz Algorithm with Exponential Convergence," *J. Fourier Anal. Appl.*, vol. 15, no. 2, pp. 262–278, Apr. 2008.

[196] A. Suárez and L. González, "Applied Mathematics and Computation," *Appl. Math. Comput.*, vol. 216, no. 2, pp. 514–522, Mar. 2010.

[197] M. J. Taghizadeh, R. Parhizkar, P. N. Garner, H. Bourlard, and A. Asaei, "Ad Hoc Microphone Array Calibration: Euclidean Distance Matrix Completion Algorithm and Theoretical Guarantees," *Signal Process.*, vol. 107, pp. 123–140, Feb. 2015.

[198] Y. Takane, F. W. Young, and J. Leeuw, "Nonmetric Individual Differences Multidimensional Scaling: an Alternating Least Squares Method with Optimal Scaling Features," *Psychometrika*, vol. 42, no. 1, pp. 7–67, Mar. 1977.

[199] I. J. Tashev, *Sound Capture and Processing*, Practical Approaches. Chichester, UK: John Wiley & Sons, July 2009.

[200] S. Tervo and T. Tossavainen, "3D Room Geometry Estimation From Measured Impulse Responses," in Proc. *IEEE ICASSP*, pp. 513–516. IEEE, Mar. 2012.

[201] M. R. P. Thomas, N. D. Gaubitch, and P. A. Naylor, "Application of Channel Shortening to Acoustic Channel Equalization in the Presence of Noise and Estimation Error," in Proc. *IEEE WASPAA*, pp. 113–116. New Paltz, NY: IEEE, 2011.

[202] S. Thrun, "Affine Structure From Sound," in Proc. *Conf. Neural Inf. Process. Sys. (NIPS).* Cambridge, MA: MIT Press, 2005.

[203] M. Toda, "Cylindrical PVDF Film Transmitters and Receivers for Air Ultrasound," *IEEE Trans. Ultrason., Ferroelect., Freq. Control*, vol. 49, no. 5, pp. 626–634, May 2002.

[204] A. Tolstoy, *Matched Field Processing for Underwater Acoustics.* World Scientific Publishing Company Incorporated, 1993.

[205] C. Tomasi and T. Kanade, "Shape and Motion From Image Streams Under Orthography: a Factorization Method," *Int. J. Comput. Vision*, vol. 9, no. 2, pp. 137–154, 1992.

[206] W. S. Torgerson, "Multidimensional scaling: I. Theory and method," *Psychometrika*, vol. 17, no. 4, pp. 401–419, Dec. 1952.

[207] J. D. Tournier, F. Calamante, D. G. Gadian, and A. Connelly, "Direct Estimation of the Fiber Orientation Density Function from Diffusion-Weighted MRI Data Using Spherical Deconvolution," *NeuroImage*, vol. 23, no. 3, pp. 1176–1185, Nov. 2004.

[208] J. A. Tropp, "Algorithms for Simultaneous Sparse Approximation. Part II: Convex Relaxation," *Signal Process.*, vol. 86, no. 3, pp. 589–602, Mar. 2006.

[209] J. A. Tropp and A. C. Gilbert, “Signal Recovery From Random Measurements Via Orthogonal Matching Pursuit,” *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.

[210] B. D. Van Veen and K. M. Buckley, “Beamforming: A Versatile Approach to Spatial Filtering,” *IEEE ASSP Mag.*, vol. 5, no. 2, pp. 4–24, 1988.

[211] F. Vanpoucke, M. Moonen, and Y. Berthoumieu, “An Efficient Subspace Algorithm for 2-D Harmonic Retrieval,” in Proc. *IEEE ICASSP*, pp. 461–464. Adelaide, SA, Australia: IEEE, 1994.

[212] D. A. Varshalovich, A. N. Moskalev, and V. K. Khersonskii, *Quantum Theory of Angular Momentum.* Singapore: World Scientific, 1989.

[213] I. N. Vekua, “New Methods for Solving Elliptic Equations,” North-Holland Publishing Company, Amsterdam, 1967.

[214] M. Vetterli, J. Kovačević, and V. K. Goyal, *Foundations of Signal Processing.* Cambridge, UK: Cambridge University Press, 2014.

[215] M. Vetterli, P. Marziliano, and T. Blu, “Sampling Signals with Finite Rate of Innovation,” *IEEE Trans. Signal Process.*, vol. 50, no. 6, pp. 1417–1428, June 2002.

[216] J. von Neumann, “Some Matrix Inequalities and Metrization of Matrix-Space ,” *Tomsk University Review*, no. 1, pp. 286–300, 1937.

[217] J. Wang, J. Liu, and J. Ye, “Efficient Mixed-Norm Regularization: Algorithms and Safe Screening Methods,” *arXiv*, July 2013.

[218] D. B. Ward, E. A. Lehmann, R. C. S. Williamson, and A. P. I. T. on, “Particle Filtering Algorithms for Tracking an Acoustic Source in a Reverberant Environment,” *IEEE Trans. Audio, Speech, Language Process.*, vol. 11, no. 6, pp. 826–836, 2003.

[219] D. B. Ward, R. A. Kennedy, and R. C. Williamson, “Theory and Design of Broadband Sensor Arrays with Frequency Invariant Far-Field Beam Patterns,” *J. Acoust. Soc. Am.*, vol. 97, no. 2, pp. 1023–1034, Feb. 1995.

[220] K. Q. Weinberger and L. K. Saul, “Unsupervised Learning of Image Manifolds by Semidefinite Programming,” *Int. J. Comput. Vision*, vol. 70, no. 1, pp. 77–90, 2006.

[221] E. Weinstein, K. Steele, A. Agarwal, and J. Glass, “LOUD: a 1020-Node Microphone Array and Acoustic Beamformer,” in Proc. *ICSV*, Cairns, Australia, July 2007.

[222] J. Wu, K. Leinenweber, J. C. H. Spence, and M. O’Keeffe, “Ab Initio Phasing of X-Ray Powder Diffraction Pattens by Charge Flipping,” *Nature Materials*, vol. 5, no. 8, pp. 647–652, Aug. 2006.

[223] W. Zhang, E. Habets, and P. A. Naylor, “On the Use of Channel Shortening in Multichannel Acoustic System Equalization,” in Proc. *IWAENC*, Tel Aviv, 2010.

[224] K. Ziętak, “Strict Spectral Approximation of a Matrix and Some Related Problems,” *Appl. Math.*, vol. 24, no. 3, pp. 267–280, 1997.

Ivan Dokmanić

Audiovisual Communications Laboratory
Ecole Polytechnique Fédérale de Lausanne
Station 14, 1015 Lausanne
Switzerland

Tel: +41 21 69 31338
Fax: +41 21 693 4312
E-mail: ivan.dokmanic@epfl.ch
http://lcav.epfl.ch/people/ivan.dokmanic

Education

Ph.D. in Communication Sciences, expected May 4, 2015
School of Computer and Communications Sciences
Ecole Polytechnique Fédérale de Lausanne
Advisor: Prof. Martin Vetterli

B.S. / M.S. in Electrical Engineering (GPA 5.00/5.00), 2007
Faculty of Electrical Engineering and Computing
University of Zagreb

Employment

Little Endian, Zagreb (October 2009 – November 2010)
Digital Audio Effects Developer (DSP Engine for SpectrumWorx software, www.littleendian.com)

University of Zagreb, Faculty of EE&C (November 2007 – September 2010)
Teaching Assistant

MainConcept AG, Aachen (September 2006 – November 2007)
Codec Developer (Development of the MainConcept ASF Muxer)

Internships

Microsoft Research Redmond (June 2013 – August 2013)
Hardware and Algorithms for Ultrasonic Depth Imaging (supervisor Ivan Tashev)

Awards and honors

Google PhD Fellowship 2014
(Speech Technologies)

Best Student Paper Award
IEEE Int. Conference on Acoustics, Speech and Signal Processing, Prague, 2011

Young Scientist Award
Society of University Teachers, Scholars and Other Scientists, University of Zagreb, 2010

Best Graduating Student in Electrical Engineering (Industrial Electronics)
University of Zagreb, 2007

"Top Scholarship for Top Students"
Nacional Magazine, 2006

Rector's Award
University of Zagreb, 2006

Dean's Honor List
University of Zagreb, Faculty of EE&C, 2003 – 2006

1st place in the Qualifying Exam (Score 1000/1000)
University of Zagreb, Faculty of Electrical Engineering and Computing, 2002

Teaching and supervision

School of Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne (2010 – 2014)

Teaching Assistant for Mathematical Signal Processing, Statistical Signal Processing and Applications, Signal Processing: Spaces, Operators and Transforms

Supervised 1 MS thesis (Pedro Oliveira Pinheiro, "Acoustic Estimation of Room Geometry with Echolocation Insights")

Supervised 5 semester projects

Faculty of Electrical Engineering and Computing
University of Zagreb (2007 – 2010)

Teaching Assistant for Embedded Systems, Multimedia Technologies, Digital Signal Processing, Digital Signal Processing Software Design, Embedded System Design, Signals and Systems

Supervised 1 MS thesis (Gordan Kreković, "Implementation of a Fast SE(3) Convolution Algorithm")

Research interests

Inverse problems in audio and acoustics

Sensor arrays

Sparse signal processing

Signal processing for sensor networks

Inference in physical fields

Professional activity

Reviewer for IEEE Transactions on Signal Processing, IEEE Journal of Selected Topics in Signal Processing, Automatika, EUSIPCO, ICASSP

IEEE Signal Processing Society Member

Languages

English	fluent
French	intermediate
Slovenian	good
Croatian	native

Invited presentations

Semidefinite Relaxations for Flexible Microphone Calibration,
European Signal Processing Conference (EUSIPCO),
Nice, September 2015

Sampling Spherical Finite Rate of Innovation Signals, *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Brisbane, April 2015

Listening to Distances and Hearing Shapes: Inverse Problems in Room Acoustics and Beyond,
Information Theory and Applications (ITA) Workshop
San Diego, February 2015

"Hearing" the Shape of a Room, and Other Treats with Echoes,
Summer School of Science, Požega, July 2014

Tricks and treats with echoes, or how computers hear room shapes,
Echolocation Event, Durham University, June 2014

Sensing Diffusion and Diffusion Like Phenomena, *Seminar on Wireless Integration of Sensor Networks in Hybrid Architectures (WISH),* Bern, March 2012

Received funds

Co-authored a project proposal funded by Swiss Kommission für Technologie und Innovation (KTI), “Touch tomography” (269084 CHF)

Received a Google PhD Fellowship (80000$ yearly)

Received an Innogrant (a start-up seed grant for a start-up with Juri Ranieri, given by EPFL’s Innovation Center) (100000 CHF)

List of publications

Journal Papers

[21] I. Dokmanic, R. Parhizkar, J. Ranieri and M. Vetterli, “Euclidean Distance Matrices: A Short Walk Through Theory, Algorithms and Applications,” To appear in *IEEE Signal Processing Magazine*, 2015.

[20] I. Dokmanic, R. Scheibler and M. Vetterli, “Raking the Cocktail Party,” To appear in *IEEE Journal of Selected Topics in Signal Processing, Special Issue on Spatial Audio*, July 2015.

[19] I. Dokmanic, R. Parhizkar, A. Walther, Y. M. Lu and M. Vetterli, “Acoustic Echoes Reveal Room Shape,” in *Proceedings of the National Academy of Sciences,* vol. 110, num. 30, 2013.

[18] I. Dokmanic and D. Petrinovic, “Efficient Approximate Scaling of Spherical Functions in the Fourier Domain With Generalization to Hyperspheres,” in *IEEE Transactions on Signal Processing*, vol. 58, num. 11, 2010.

[17] I. Dokmanic and D. Petrinovic, “Convolution on the n-Sphere With Application to PDF Modeling,” in *IEEE Transactions on Signal Processing*, vol. 58, num. 3, 2010.

[16] I. Dokmanic, M. Sikic and S. Tomic, “Metals in Proteins: Correlation Between the Metal-Ion Type, Coordination Number and the Amino-Acid Residues Involved in the Coordination,” in *Acta Crystallographica Section D-Biological Crystallography,* vol. 64, 2008.

Submitted Journal Papers and Journal Papers In Preparation

[15] I. Dokmanic and R. Gribonval, “Beyond Moore-Penrose? Generalized Inverses that Minimize Matrix Norms,” To be submitted to *Linear Algebra and its Applications*, 2015.

[14] I. Dokmanic and Y. M. Lu, “Sampling Sparse Signals on the Sphere: Algorithms and Applications,” submitted to *IEEE Transactions on Signal Processing,* 2015.

Conference Papers

[13] I. Dokmanic and M. Vetterli, “Semidefinite Relaxations for Flexible Microphone Calibration,” Invited to *European Signal Processing (EUSIPCO),* Nice, September 2015.

[12] I. Dokmanic and Y. M. Lu, “Sampling Spherical Finite Rate of Innovation Signals,” *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Brisbane, April 2015.

[11] R. Scheibler, I. Dokmanic and M. Vetterli, “Raking Echoes in the Time Domain,” *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Brisbane, April 2015.

[10] I. Dokmanic, L. Daudet and M. Vetterli, “How to Localize Ten Microphones in One Fingersnap,” *European Signal Processing (EUSIPCO)*, Lisbon, September 2014.

[9] I. Dokmanic and I. Tashev, “Hardware and Algorithms for Ultrasonic Depth Imaging,” *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Florence, May 2014.

[8] O. Öçal, I. Dokmanic and M. Vetterli, “Source Localization and Tracking in non-Convex Rooms,” *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Florence, May 2014.

[7] R. Parhizkar, I. Dokmanic and M. Vetterli, “Single-Channel Indoor Microphone Localization,” *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Florence, May 2014.

[6] I. Dokmanic, M. Kolundzija and M. Vetterli, “Beyond Moore-Penrose: Sparse Pseudoinverse,” *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, May 2013.

[5] M. Martinez-Camara, I. Dokmanic, J. Ranieri, R. Scheibler, M. Vetterli and Andreas Stohl, “The Fukushima Inverse Problem,” *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, May 2013.

[4] I. Dokmanic and M. Vetterli, “Room Helps: Acoustic Localization With Finite Elements,” *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Kyoto, March 2012.

[3] J. Ranieri, I. Dokmanic, A. Chebira and M. Vetterli, “Sampling and Reconstruction of Time-Varying Atmospheric Emissions,” *IEEE International*

Conference on Acoustics, Speech, and Signal Processing (ICASSP), Kyoto, March 2012.

[2] I. Dokmanic, J. Ranieri, A. Chebira and M. Vetterli, “Sensor Networks for Diffusion Fields: Detection of Sources in Space and Time,” *Allerton Conference on Communication, Control, and Computing*, Monticello, September 2011.

[1] I. Dokmanic, Y. Lu and M. Vetterli, “Can One Hear the Shape of a Room: The 2-D Polygonal Case,” *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, Prague, May 2011.

List of Patents

Issued

[2] Optical Touch Tomography, November 2013.

[1] A Method and a System for Determining the Geometry and/or Localization of an Object, June 2014.

Filed

[5] Ultrasonic Depth Imaging, September 2014.

[4] Optimal Acoustic Rake Receivers, July 2014.

[3] Calibration Method and System, March 2014.

[2] A Method and a System for Determining the Location of an Object (non-convex localization), December 2013.

[1] A Method and a System for Determining the Location of an Object (single channel localization), December 2013.