UCLouvain
Institute for Information
and Communication Technologies,
Electronics and Applied Mathematics

One Bit at a Time
The Use of Quantized Compressive Sensing in RADAR Signal Processing

Thomas Feuillen

Thesis submitted in partial fulfillment
of the requirements for the degree of
Ph.D. in Engineering Sciences

Dissertation committee:

Prof. Laurent Jacques (UCLouvain, advisor)
Prof. Luc Vandendorpe (UCLouvain, advisor)
Prof. Christophe Craeye (UCLouvain)
Prof. Mike E. Davies (Edinburgh University, Scotland)
Dr. Matthias Weiß (Fraunhofer FHR, Germany)
Prof. Laurent Francis (UCLouvain)

Version of October 2021.

Contents

List of Figures

Abstract

THIS thesis studies the harsh quantization of radar signals. More specifically, what can be achieved in terms of localization of targets using FMCW radars from 1-bit dithered measurements and processing. The first part of this thesis leverages the framework of Quantized Compressive Sensing to achieve high quality localizations using coarse 1-bit measurements from an FMCW radar. The gain provided by the added dither is highlighted through simulations and actual radar measurements and are compared with the developed reconstruction bounds. Range and angle estimations are achieved using the PBP and QIHT algorithms. The second part highlights some difficulties inherent to adding a random dither to radar signals and in response, studies an alternative way of dithering the measurements by altering instead their phases. This method, compared to the additive case, is shown to be a viable alternative in the search for an implementation of 1-bit quantization of radar signal that has theoretical guarantees and is cost-effective to implement. This new way of dithering radar signals is compared using Monte-Carlo simulations against its additive counter-part and using actual radar data. This alternative way of dithering is linked to the Phase-Only acquisition, that only measures the phase of complex signals, and its reconstruction performances are studied through the lens of the guarantees provided to PBP using the (ℓ_1, ℓ_2)-Restricted Isometry Property. This property is proved for complex Gaussian random matrices. The thesis does not finish by the study of yet another way of acquiring a quantized version of a signal but by studying the quantization of the processing itself. Indeed, using low resolution processing could enable more power-efficient implementations. To that end, we study the reconstruction guarantees of the Projected Back Projection algorithm in

the setting where the back-projection used is a 1-bit quantized version with additive dithering of the one used in high resolution processing. We show a uniform bound on the ℓ_2 reconstruction that behaves as $\mathcal{O}(m^{-\frac{1}{2}})$. This study is then extended to the case of back-projection operators that have a factorized representation. These factorized representations, among which the FFT is the most well-known, can often be computed efficiently thanks to their sparse and factorized structures.

This thesis shows that in cases where either the power or the amount of data that one can use for the estimation is limited, lowering the individual resolution of the measurements and possibly of the processing, can allow for better results than sub-sampling those high-resolution measurements to fit within the limitations. This was shown throughout the thesis using both theory and simulations often accompanied by real radar measurements.

Acknowledgements

In the first year of my PhD I remember one other PhD student, who was in the process of writing his thesis, telling me: "*You'll see, a PhD is not a race but a marathon*". As I am now seeing the proverbial finish line, it is time to reflect on the path taken and those who helped me along the way.

First, I would like first to thank my two supervisors Laurent Jacques and Luc Vandendorpe. This thesis could not have been possible if not for them. Laurent is one of the best PhD supervisor that anyone can have. His motivation, insights, patience and empathy, helped me throughout my PhD. Apart from its scientific supervision, Laurent also helped me foster collaborations and gave me numerous opportunities to present my work to other people in the field. I am really grateful to have been a part of his lab.

I would also like to thank Luc Vandendorpe. He was the one who, at the end of my master thesis, suggested that I could do a PhD and helped me become a teaching assistant. Luc was instrumental to me becoming the researcher that I am today and helping me from the early stages of my research until the very end. From him I also learned how to teach and create projects that challenged but also rewarded students. He also always made sure that I had access to any systems that I needed to conduct my research and that I had the necessary means to attend numerous doctoral schools, workshops and conferences. In hindsight, I realise that this is not a chance that all PhD students are afforded, and for that I am grateful.

I would also like to thank the jury members of my thesis, Prof Christophe Craeye, Prof. Mike Davies, Dr. Matthias Weiß and Prof. Laurent Francis for reviewing this thesis and for their comments. I started my research in radars with Prof. Christophe Craeye who was the co-supervisor of my

master thesis with Luc Vandendorpe. Christophe Craeye has always been there to help me in all the stages of my PhD. He is a great researcher but also a great motivator and someone you can always count on. He is also able to create a real team spirit within his group and to make everyone feel welcome.

During my scientific stay in Edinburgh, I had the opportunity to join the team of Mike Davies. These three months spent in his team are one of the highlights of my PhD. I really enjoyed working alongside him and discovering Scotland.

Matthias Weiß was one of the few familiar faces that I would see at conferences and workshops on radar signal processing. He also organised the COSERA conference, where I first presented my work on quantized compressive sensing applied to radar, and he is also one of the main organizer of the famous Radar Summerschool. Both were great events that I enjoyed thanks to his organization.

I would also like to thank three colleagues that tremendously helped me during my PhD. Chunlei Xu, who was a post-doc and helped me to enter the sometimes opaque mathematical realm that underpins the use of dithering in quantized compressive sensing. She helped quickstart my scientifical research in the application of this theory to radar signal processing. Although he would deny it, I could not have gone this far without Amirafshar Moshtaghpour's help and advice. He is a great colleague and a dear friend. Finally, I would like to thank Alexander Stollenwerk whose comments and insights helped in the writing of the last part of this manuscript.

I would also like to thank all the friends made during these years in ICTEAM : Thomas Pairon, Maxime Drouguet, Amirafshar Moshtaghpour, Charles Wiame, Jean Cavillot, Martin Delcourt, Gilles Monnoyer, Julian David Villegas Gutierrez, Husnain Ali Kayani, Hussein Kassab, Valerio Cambareri, Domingo Pimienta del Valle, Nafiseh Janatian, Chunlei Xu, Souley Djadjandi, Antoine Paris, Jean Leger, Mohieddine El Soussi, Adriana Gonzalez, Pierre Gerard, Brigitte Dupont, Francois Hubin, Claude Oestges, Alexander Stollenwerk, Pascal Simon, Simon Demey and Isabelle Dargent.

I would also like to thank Callum, Mario and Mikey, who always made me feel welcome in their lab in Edinburgh and showed me the best pubs in the city, where unforgettable memories were made.

My friends who supported me in the good and bad times during my PhD : Remi, Martin, Khoi, Jean-Em, Alan, Damien, Jerome, Thibaud, Mika, Mauro, Denis, David, Hugo, Nicolas.

I would also like to thank my family for their support.

Finally, without the unwavering support and bottomless patience of my girlfriend Cindy, this manuscript would be nowhere from finished.

Abbreviations & Notations

Abbreviations

ADC	Analog to Digital Converter
BP	Back-Projection
BPDN	Basis-Pursuit-DeNoising
CS	Compressive Sensing
CW	Continuous Wave
FFT	Fast Fourier Transform
FMCW	Frequency Modulated Continuous Wave
FPGA	Field Programmable Gate Arrays
IHT	Iterative Hard-Thresholding
LIDAR	Light Detection And Ranging
MRI	Magnetic Resonance Imaging
PBP	Projected-Back-Projection
PO	Phase-Only
PRF	Pulse Repetition Frequency

QIHT	Quantized Iterative Hard-Thresholding
QPBP(Q)	Quantized Projected-Back-Projection of (possibly Quantized) measurements
RCS	Radar Cross Section
RF	Radio Frequency
RIP	Restricted Isometry Property

Notations

$(\cdot)^*$	denotes the complex conjugate and the adjoint operator for scalar and matrices, respectively.
$[D]$	$[D] := \{1, \cdots, D\}$ for $D \in \mathbb{N}$
$\angle(\cdot)$	$\angle(re^{\mathsf{i}\,\alpha}) = \alpha$
$\mathbb{B}^N$	The ℓ_2 (or Frobenius) unit ball in $\mathbb{R}^N$ (resp. $\mathbb{R}^{D\times D'}$) is denoted by $\mathbb{B}^N$ (resp. $\mathbb{B}_F^{D\times D'} \simeq \mathbb{B}^{DD'}$).
$\mathbb{Z}_\delta$	$\mathbb{Z}_\delta := \delta\mathbb{Z} + \delta/2$
$\boldsymbol{B}_{:,j}$ (or $\boldsymbol{B}_{j,:}$)	are the j^{th} column (resp. row) of $\boldsymbol{B}$.
$\boldsymbol{B}_{\mathcal{S}}$, $\boldsymbol{u}_{\mathcal{S}}$	For any $\boldsymbol{B} \in \mathbb{C}^{D\times D'}$ (or $\boldsymbol{u} \in \mathbb{C}^{D'}$), $\boldsymbol{B}_{\mathcal{S}}$ (resp. $\boldsymbol{u}_{\mathcal{S}}$) is the cropped matrix (resp. vector) obtained by restricting the columns (resp. components) of $\boldsymbol{B}$ (resp. $\boldsymbol{u}$) to those indexed in $\mathcal{S} \subset [D']$
$\boldsymbol{x}, \boldsymbol{\Phi}$	Vectors and matrices are denoted with bold symbols
$\mathcal{U}_\delta^{\mathbb{C}}$	its complex counterpart is $\mathcal{U}_\delta^{\mathbb{C}} := \mathcal{U}_\delta^{\mathbb{R}} + \mathsf{i}\,\mathcal{U}_\delta^{\mathbb{R}}$
$\mathcal{U}_\delta^{\mathbb{R}}$	The uniform distribution over $[-\frac{\delta}{2}, \frac{\delta}{2}]$ is denoted $\mathcal{U}_\delta^{\mathbb{R}}$
$\Gamma_T(t)$	the window function of length T
i	The imaginary unit is $\mathsf{i} = \sqrt{-1}$
$\lfloor\cdot\rfloor$	is the flooring operator

$\|\boldsymbol{B}\|_F$	the Frobenius norm and scalar product of matrices are related by $\|\boldsymbol{B}\|_F = (\langle \boldsymbol{B}, \boldsymbol{B} \rangle_F)^{1/2} = (\mathrm{tr}\, \boldsymbol{B}^* \boldsymbol{B})^{1/2}$
$\|\boldsymbol{u}\|_0$	ℓ_0-norm of $\boldsymbol{u}$ defined as $\|\boldsymbol{u}\|_0 =\mid \mathrm{supp}(\boldsymbol{u}) \mid$
$\|\boldsymbol{u}\|_p$	For $p \geq 1$, the ℓ_p-norm of a complex vector $\boldsymbol{u}$ reads $\|\boldsymbol{u}\|_p := (\sum_k \mid u_k \mid^p)^{1/p}$, with $\|\boldsymbol{u}\| := \|\boldsymbol{u}\|_2$ and $\|\boldsymbol{u}\|_\infty = \max_k \mid u_k \mid$.
$\mathbf{1}_N$	is the unitary vector of size d that made only of 1.
c	the speed of light
$\mid \mathcal{S} \mid$	is the cardinality of a set $\mathcal{S}$
$\mathrm{supp}(\boldsymbol{u})$	$\mathrm{supp}(\boldsymbol{u}) = \{i : u_i \neq 0\}$ is the support of $\boldsymbol{u}$
$B_{\mathbb{R}}$, B_{I}	for any complex quantity B, *e.g.*, a scalar, a vector or a matrix, $B_{\mathrm{R}} = \Re(B)$ and $B_{\mathrm{I}} = \Im(B)$ are the real and imaginary parts of B, respectively
$x \bmod y$	is the modulo operator of size y applied to x
$\mathbf{Id}$	is the identity matrix
$\bar{\mathbb{B}}^N$	is the ℓ_2 unit ball in $\mathbb{C}^N$
Σ_s^N	is the set of real s-sparse vectors in $\mathbb{R}^N$
$\bar{\Sigma}_s^N$	is the set of complex s-sparse vectors in $\mathbb{C}^N$
$\tilde{\Sigma}_s^N$	is the set of complex s-sparse vectors in $\bar{\mathbb{B}}^N$

Part I

Introduction

Chapter 1

Introduction

Since the first expirements made by Christian Hülsmeyer in 1904 in Cologne, radars have continued to grow in popularity. From being only able to detect the presence of boats to being able to detect the effect of climate change from measurements performed from satellites in space, radar technology has always focused on remotely detecting and measuring its environment. In previous decades, radar technology was strictly confined to applications managed by government agencies such as remote sensing research, military and the police. Recently, radars have now conquered new markets and will be more and more omnipresent in day-to-day use, from autonomous cars, to smart cities, home automation and cellphones.

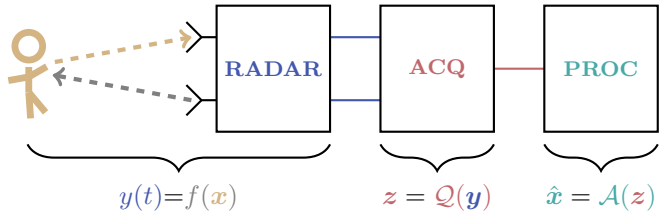


Figure 1.1: Representation of the different processes involved in radar detection.

In classic radar signal processing, the **radar** transmits a *high frequency* signal that interacts with its **environment** and is reflected back to the radar. The radar then demodulates this signal and generates a *low frequency* analogic signal $y(t)$. To be processed, this signal first needs to be sampled by

Analog to Digital Converters (**ADCs**). These ADCs record the signal $y(t)$ in the digital domain according to a sampling frequency and a resolution *i.e.*, $z[n] = \mathcal{Q}(y(nT_s))$. This **acquisition** process is called *quantization*. The digital measurements are then transmitted to the **Processing Unit** where an algorithm estimates the desired properties of the environment $\tilde{x}$ from the quantized measurements z.

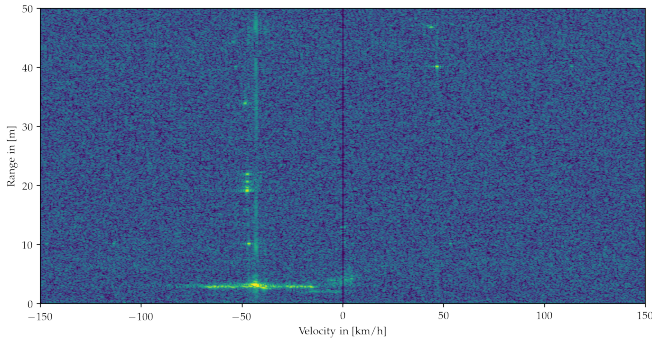
Because the signals are digitized to a finite number of bits, the quantization $\mathcal{Q}(y(t))$ cannot capture perfectly the analogical signal $y(t)$, inducing a discrepancy between the two. In classic radar signal processing, the effect of the ADCs on the measurements is often omitted. To that end, $y(t)$ is sampled according to the *Shannon*'s theorem and the *resolution* and the *dynamic* of the ADCs are extremely high so that the quantization is negligible, *i.e.*, $z \approx y$.

This way of acquiring data is really effective in applications where neither the cost of the hardware, the amount of data to be transmitted between the acquisition and the processing unit, nor the power required are limited. But in new radar applications that are more constrained, this represents a challenge because of their size and cost.

On the one hand, new radars are often *Multi-Input Multi-Output*, which means radars that have more than one transmitting and receiving antennas, which multiplies the amount of data that needs to be recorded. Having high resolutions ADCs might create too much data that need to be transmitted and processed in a timely manner. On the other hand, fast and high resolution ADCs have a certain cost and power requirements that are not well suited to these new applications.

One way of easing these requirements is to lower the number of measurements required, *i.e.*, going below *Nyquist*'s rate. Among the numerous methods that try to lower the amount of measurement required to achieve high quality of reconstruction, *Compressive Sensing* is of particular note. The *Compressive Sensing* (CS) theory leverages the low-complexity nature of structured signals (*e.g.*, their sparsity, compressibility or low-rankness) to reduce the signal sampling rate at the acquisition [CRT06b; FR13]. In a radar context, scenes that are illuminated by a radar can often be represented as sparse. For example Fig. 1.2b is the *Range-Doppler* map obtained using a *Frequency Modulated Continuous Wave* (FMCW) radar from the

(a)



(b)

Figure 1.2: (a) Picture of a scene that was measured with the KMD2 radar by [RFB]. (b) The corresponding Range-Doppler obtained by a 2D-FFT on the received signal.

measurements collected from the scene in Fig. 1.2a. The signal transmitted by the radar is reflected by all the objects present in the scene. Metallic objects, however, reflect the emitted power at orders of magnitude above other materials such as wood [Sko80] therefore the *Range-Doppler* map has only a few high amplitude points corresponding to the different cars in Fig. 1.2a, making the observed scene effectively sparse. CS shows that, with high probability, one can stably and robustly estimate such signals by collecting a number of random linear measurements driven by the signal "information-rate", *e.g.*, its sparsity level (which in this example corresponds to the number of cars). During the last ten years, many works have considered the association of the radar principles with CS theory: first, to increase a tar-

get's parameter resolution [HS09], and later to reduce the number of samples to be processed [End13]. The survey [CE18] describes the reduced sampling rate of different compressive (or sub-Nyquist) radar systems, in comparison with traditional Nyquist sampling schemes, although digitization impact is not covered.

The classical CS setting still considers that the measured signals are of infinitely high resolution, which in turns requires the use of costly and high power ADCs. In the first part of this thesis, we propose to remove this limitation by integrating the quantization directly in the signal model, using the framework of *Quantized Compressive Sensing* (QCS). More specifically, we focus on lightening the acquisition of radar signals by strongly lowering the resolution (or bit-depth) of each sample collected by a radar sensor without sacrificing accurate depth estimation. We consider several different quantization procedures and study both their theoretical guarantees and their applicability in the context of radar signal processing. A special interest will be taken to the harsh 1-bit quantization with dithering of these radar signals, effectively only recording the signs of their analog signal to which a random variable is added.

The second part of this thesis focuses not only on lightening the acquisition but also on the processing that estimates the signal of interest from possibly quantized measurements. Indeed, regardless of the resolution of the individual measurements, the algorithms used for the reconstruction are performed using methods that are often designed for high-resolution signals. Such high-resolution processing requires complex hardware architecture with high power demand. Similarly to the quantization noise generated by coarse acquisitions, lowering the resolution of the processing will also impact the quality of the reconstruction. In the last part of the thesis, we study how to efficiently lower the resolution of the Back-Projection to 1-bit and study its effect on the reconstruction.

1.1 Contributions

What this thesis is not is a search for the best reconstruction performances of complex radar scenes, regardless of theoretical guarantees. It is rather an exploratory venture into the highly theoretical field of Quantized Compres-

sive Sensing through the lens of radar signal processing. Consequently, the algorithms used in this thesis do not provide the best reconstructions but attempt to offer the strongest guarantees while remaining consistent with actual radar applications. This work tries to combine the highly theoretical field and results from Quantized Compressive Sensing with the more applied setting of FMCW radar signal processing. To that end, we deliberately chose a restrictive and idealized radar setting where, for example, the measurements are noiseless, and explore what can be shown theoretically about these scenarios and then demonstrate these conclusions practically using real radar measurements. This restrictive setting allows us to highlight specific effects imparted on the signal by the 1-bit quantization that a more complex scenario with, for example noise and clutter, would overshadow. This work thus focuses on the careful interplay between theoretical considerations and their real world impact in radar signal processing.

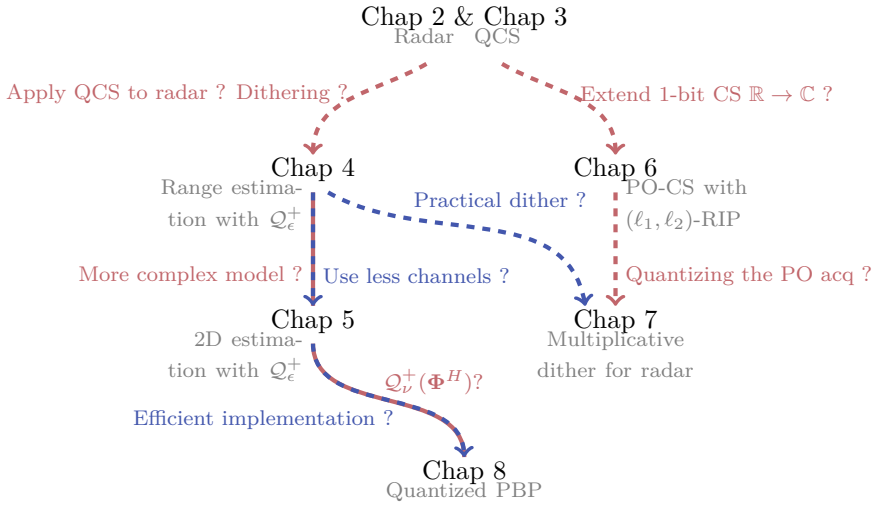


Figure 1.3: Link between the chapters of the thesis; Chapters linked through the Theory in **red** (as a direct applications or extensions); Chapters linked by practical considerations in **blue**.

In Fig. 1.3, we present the different connections and interactions between the different chapters of this thesis. Chapters 2 and 3 constitute a non-exhaustive summary of FMCW radars and (Q)CS. These chapters provide

the necessary background and set the stage for the study of quantization schemes applied to radar.

In Part II, the harsh quantization to 1-bit of radar signals coming from an FMCW radar is studied. We show that in order to have successful reconstructions of sparse scenes where the resolution of the acquisition has been lowered to 1-bit, one must add a random dithering before the quantization.

- In chapter 4, we compare different resolutions of ADCs (namely one-bit, 2-bits and high resolution linear measurements) using the common metric of the bit-rate, *i.e.*, the number of bits that are stored to represent the acquired data used for the reconstruction. We show that lowering dramatically the resolution to 1-bit, which allows one to increase the number of measurements for a given bit-rate, gives better reconstruction results than using a low number of high quality measurements; in other words *quantity over quality*. Furthermore, we show that in the case of Fourier based measurements, a direct 1-bit quantization results in possible ambiguous scenarios that cannot be estimated accurately without the addition of a dither. These results are highlighted using extensive Monte-Carlo simulations in the setting of reconstruction the range profile observed by the radar. The results from the simulations are then confirmed using real radar measurements thanks to a hardware device that is able to simulate a scene of targets in front of an actual radar. This experimental set-up shows that adding a random dither before the quantization is a necessity in practical settings for the reconstruction to be successful.

- In chapter 5, we extend the sparse recovery problem to a 2-D setting where the FMCW radar has two receiving antennas. We show first that, similarly to the previous 1-D ranging problem, estimating the angle of arrival from the received 1-bit measurements can be a highly ambiguous process if no dither is added before the quantization, especially when only one target is measured. This is shown theoretically and confirmed through simulations. Second, we show that the amount of data required to perform the localization could even be further reduced by omitting the real or imaginary part of the measured signal on the different antennas, halving the number of measurements and the number of ADCs required whithout sacrificing on the maximum

estimated range. Lowering the resolution to 1-bit combined with this *channel dropping* generated a gain of almost 98% of data compression in terms of allocated bit rate compared to the classic acquisition scheme.

In Part III, an alternative way of lowering the resolution of signals is studied. Indeed, while adding a random dither has the benefit of strong theoretical guarantees when used in conjunction with the PBP algorithm, implementing a cost efficient additive dithering is a rather challenging task. We thus study alternative and cost efficient ways of lowering the resolution of radar measurements.

- In chapter 6, the coarse Phase-Only acquisition is studied (*i.e.*, $\mathrm{sign}_{\mathbb{C}}$). This acquisition only measures the phase of the complex measurements and discards the amplitude. We show that provided that the measurement matrix follows the (ℓ_1, ℓ_2)-RIP $(2s, \delta)$, then the reconstruction obtained by the PBP algorithm can be bounded by the RIP constant δ. We further show that the complex Gaussian measurement follows this property for a number of measurements m sufficiently high. Although this acquisition process is a bit more removed from any direct application such as radar or others, it offers insights into other acquisition processes that quantize, for example, the phase of complex signals.

- In chapter 7, we propose to multiplicatively dither the measurements, *i.e.*, dithering their phases. We show that the advantages of this dithering process are threefold: (*i*) this dithering procedure can be efficiently implemented in FMCW radar architecture using *off-the-shelf* components, (*ii*) using PBP algorithm, the multiplicative dithering procedure is more resistant to high sparsity signals, (*iii*) the constraint on the random phase of the dither can be relaxed to a deterministic single tone complex exponential which can be efficiently implemented in hardware. These advantages, however, come at a cost in terms of theoretical guarantees. Indeed, we show that this 1-bit quantization with a dithered phase can be related to the Phase-Only acquisition and we introduce, similarly to Chapter 4 and Chapter 5, ambiguous scenarios that demonstrate the impossibility of uniform guarantees

using Fourier based measurements. However, we show that the discrepancy between the reconstruction obtained using proposed 1-bit quantization and the PO measurements can be upper-bounded using a non-uniform proof. The simulations and real radar measurements show that this way of dithering the measurements represents an interesting trade-off between performances, complexity of implementation and theoretical guarantees.

In the third and last part, we shift the focus from the quantization of measurements to the quantization of the reconstruction process itself.

- In chapter 8, we study the quantization of the Projected-Back-Projection algorithm (QPBP). We first develop uniform guarantees on the ℓ_2-reconstruction when a 1-bit equivalent of the measurement matrix is used for the reconstruction. This shows that the reconstruction error can be made arbitrary low for a sufficient number of measurements, regardless of their resolution. In the second part of this chapter, we extend the study to back-projection operators that can be factorized into multiple sub-matrices whose lines are sparse and have a fast matrix-vector multiplication. One example of this type of back-projection is the ubiquitous Fast Fourier Transform. Again we develop uniform recovery guarantee for linear and 1-bit measurements (*i.e.*, QPBP and QPBPQ). These bounds are then assessed using Monte-Carlo simulations.

A summary of these contributions and how they interact with the state of the art is presented in Table 1.1.

1.2 List of Publications

1.2.1 Journal Papers

- (To Be Submitted) **T. Feuillen**, A. Stollenwerk, L. Vandendorpe, L. Jacques, *One Bit to Rule them All, Quantizing the Backprojection for Factorizable Models*; IEEE Transactions on Signal Processing
- (To Be Submitted) **T. Feuillen**, L. Vandendorpe, L. Jacques, *Multiplicative Dithering for 1-Bit Radar Signal Pocessing*; IEEE Transac-

	Sensing Method			
Algorithm	$\boldsymbol{\Phi x}$	$\mathcal{Q}_\epsilon^+(\boldsymbol{\Phi x})$	$\mathcal{Q}^\odot(\boldsymbol{\Phi x})$	$\text{sign}_\mathbb{C}(\boldsymbol{\Phi x})$
PBP	Classic	Applied to Radar	Non-uniform reconstruction bound, applied to radar	Uniform reconstruction, (ℓ_1, ℓ_2) RIP for complex gaussian
IHT	Classic	Applied to Radar	Applied to radar	Applied to radar
QPBP	Uniform reconstruction bound	Uniform reconstruction bound		

Table 1.1: Table representing part of the contributions of this thesis; The cells in **blue** are subjects already covered in other works; **green** cells are the subjects of this thesis; the **red** cells are left as open questions and possible future works.

tions on Aerospace and Electronic Systems

- (2020, Submitted) H. Kassab, F. Rottenberg, **T. Feuillen**, C. Wiame, J. Louveaux, *Superposition of Rectangular Power Pulses and CP-OFDM Signal for SWIPT*; EURASIP Journal on Wireless Communications and Networking

- (2020) D. Dardari, N. Decarli, A. Guerra, M. Fantuzzi, D. Masotti, A. Costanzo, D. Fabbri, A. Romani, M. Drouguet, **T. Feuillen**, C. Raucy, L. Vandendorpe, C. Craeye, *An Ultra-Low Power Ultra-Wide Bandwidth Positioning System*; IEEE Journal of Radio Frequency Identification 4 (4), 353-364

- (2020) **T. Feuillen**, M.E. Davies, L. Vandendorpe, L. Jacques, *(ℓ_1, ℓ_2)-RIP and Projected Back-Projection Reconstruction for Phase-Only Measurements*; IEEE Signal Processing Letters 27, 396-400

- (2020) L. Jacques, **T. Feuillen**, *The importance of phase in complex compressive sensing*; ArXiv preprint 2001.02529; IEEE Transactions on Information Theory, doi: 10.1109/TIT.2021.3073566.

- (2017) **T. Feuillen**, T. Pairon, C. Craeye, L. Vandendorpe, *Localization of Rotating Targets Using a Monochromatic Continuous-Wave Radar*; IEEE Antennas and Wireless Propagation Letters 16, 2598-2601

1.2.2 Conference Papers

- (2021) G. Monnoyer, **T. Feuillen**, L. Vandendorpe, L. Jacques, *Sparse Factorization-based Detection of Off-the-Grid Moving targets using FMCW radars*; ArXiv 2102.05072, ICASSP 2021

- (2020) L. Jacques, **T. Feuillen**, *Keep the phase! Signal recovery in phase-only compressive sensing*; ArXiv 2011.06499; iTWIST'20.

- (2020) G. Monnoyer, **T. Feuillen**, L. Vandendorpe, L. Jacques, *Going Below and Beyond, Off-the-Grid Velocity Estimation from 1-bit Radar Measurements*; ArXiv 2011.05034, RadarConf 2021.

- (2020) **T. Feuillen**, M.E. Davies, L. Vandendorpe, L. Jacques, *One Bit to Rule Them All: Binarizing the Reconstruction in 1-bit Compressive Sensing*; ArXiv 2008.07264; iTWIST'20.

- (2019) G. Monnoyer, **T. Feuillen**, L. Jacques, L. Vandendorpe, *Sparsity-driven moving target detection in distributed multistatic FMCW radars*; 2019 2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP).

- (2019) D. Dardari, N. Decarli, D. Fabbri, A. Guerra, M. Fantuzzi, D. Masotti, A. Costanzo, A. Romani, M. Drouguet, **T. Feuillen**, C. Raucy, L. Vandendorpe, C. Craeye, *An Ultra-wideband Battery-less Positioning System for Space Applications*; 2019 IEEE International Conference on RFID Technology and Applications (RFID-TA), 104-109.

- (2019) **T. Feuillen**, C. Xu, J. Louveaux, L. Vandendorpe, L. Jacques *Quantity over quality: dithered quantization for compressive radar systems*; 2019 IEEE Radar Conference (RadarConf), 1-6.

- (2019) **T. Feuillen**, L. Vandendorpe, L. Jacques, *An extreme bit-rate reduction scheme for 2D radar localization*; ArXiv 1812.05359; iTWIST'18.

- (2018) **T. Feuillen**, C. Xu, L. Vandendorpe, L. Jacques, *1-bit Localization Scheme for Radar using Dithered Quantized Compressed*

Sensing; ArXiv 1806.05408; 2018 5th International Workshop on Compressed Sensing applied to Radar, Multimodal Sensing, and Imaging (CoSeRa).

- (2016) **T. Feuillen**, A. Mallat, L. Vandendorpe, *Stepped frequency radar for automotive application: Range-Doppler coupling and distortions analysis*; MILCOM 2016-2016 IEEE Military Communications Conference, 894-899

The results in [Feu+19; Feu+18a; Feu+18b; FVJ18; Feu+20] are presented within the different chapters of this thesis along with the results of the soon to be submitted papers. The publications [FMV16; JF21; Dar+19; Dar+20; Feu+17; Gal+19; Gal+20; Gal+21] are not included for conciseness as their contributions are more tangential to the topic of the thesis. A short summary of these papers is provided hereafter. In [FMV16], the Range-Doppler coupling that arises in applications with high velocity targets is tackled using a computationally efficient filter. In [Feu+17], a cost effective method of estimating the range of rotating targets using a single tone Doppler radar is proposed. This method relies on the structured spectral signature estimated from the received signal to infer the range of a rotating target. These results were confirmed using real radar measurements. Passive tag localisation using Ultra Wide Band technology is studied in [Dar+19; Dar+20]. These works were conducted within the framework of an European Space Agency (ESA) project in collaboration with the University of Bologna in Italy, and culminated in a live demonstration in the ESA headquarters of the UWB battery-less localization system. My contribution to these works was to develop a computationally efficient correlation of the UWB signals used in a time-of-arrival estimation scheme. This method relies on the structure of the spectrum of the pulse that allowed an undersampling of the received signal that is then leveraged in order to compute the correlation. The system showcased an accuracy of $\sim$ 4cm. The works presented in [Gal+19; Gal+20; Gal+21] are a joint work with Gilles Monnoyer, a PhD student supervised by Laurent Jacques and Luc Vandendorpe in UCLouvain whose master theses I also supervised. These works focus on the efficient estimation of the *off-the-grid* target's parameter such as the position and velocity in applications using MIMO radars by relying on the factorization of the signal's representation in order to reach efficient computations. This *off-the-grid*

estimation scheme was combined in [Gal+20] with a one-bit quantization scheme with dithering that will be introduced in Chapter 3. The letter [Feu+19] that is the focus of Chapter 6 was extended in [JF21], where a non-uniform reconstruction guarantee is studied for the reconstruction of sparse vectors from Phase-Only measurements using *Instance Optimal* algorithm such as Basis Pursuit Denoising. In *Superposition of Rectangular Power Pulses and CP-OFDM Signal for SWIPT*, that has been submitted to the EURASIP Journal on Wireless Communications and Networking, a new *SWIPT* modulation waveform is introduced. My contribution to this work was in the test-bed used for practical measurements with USRPs. Although not directly connected to the field of quantized compressive sensing, all of these articles are centred around the idea of implementing cost or energy efficient methods to estimate parameters from incomplete or imperfect measurements.

Finally, during my thesis I also had the opportunity to supervise several master thesis. For the sake of conciseness, these are only listed hereafter. *Radar target classification based on micro-Doppler signature analysis* in 2016 by Jean Leger; *Synthetic aperture radar at small scale* in 2019 by Marie-Pierre van Oldeneel and Adrien Delhaye; *Sparsity-driven moving target detection in distributed multistatic FMCW radars* in 2019 by Gilles Monnoyer; *Analysis of a channel of communication for on-board Drone transmission with SDR processing* in 2020 by Dylan Feron; *Formation of high resolution images via Synthetic Aperture Radar embedded on satellite* in 2020 by Cyril Wain.

Chapter 2

FMCW Radar System

In this chapter, a short review of the model and architecture of Frequency Modulated Continuous Wave (FMCW) radars are presented. We start by a short history of radar technology and its applications before reviewing the basic mode of operation and limitation of a pulse radar. The next section is then devoted to the FMCW radar. The signal model corresponding to the observation of a scene with multiple targets located at different ranges is presented as well as common non-idealities typically found in such radar systems.

2.1 Introduction

Radar stands for *Radio Dectection And Ranging*. It is a device that transmits electromagnetic waves through an antenna and then uses the received echoes to infer information about the scene it has interacted with.

There is no definite sole inventor of the radar in the modern sense of the word as it was independently studied by various countries during World War 2 [Sko80]. However, one inventor in 1904 can be credited with the first wireless detection of metallic object using electromagnetic waves [Hul06]. Christian Hülsmeyer demonstrated in Cologne his *telemobiloscope* in front of German military officials but his invention failed to gain traction. One reason was that this early prototype was only able to detect the presence of ships but not their range, furthermore the detection range was of only a

few hundred meters [Sko80].

During WW2, the ability to detect enemy ships and planes became of paramount importance. Consequently radar technology evolved at a rapid pace during this period, from the ability to detect targets from a few miles to several tens of miles. Although Hülsmeyer *telemobiliscope* used pulses, early radar research focused first on Continuous Wave modulation before switching to pulses. These changes of modulations were dictated by the advancements in the *Radio-Frequency* hardware that these early radars were using.

Regardless of the modulation used, radars are governed by the following equation:

$$P_r = \frac{P_t G A \sigma}{(4\pi)^2 R^4}, \tag{2.1}$$

where P_r is the power density of the signal coming from a target at a range R, G is the gain of the transmitting antenna, A is the effective area of the receiving antenna and σ is the *Radar-Cross-Section* (RCS).

The gain of the transmitting antenna in the direction of the target is defined with respect to the gain of an isotropic antenna that radiates the same power in all direction. Depending on the application, it is desirable, or not, to have antennas that are highly directive, *i.e.*, that can transmit and receive power in specific directions. The amount of power reflected back to the radar by the target is characterized using an area called the RCS. It is equivalent to the cross-sectional area of a perfectly reflecting sphere that would radiate the same power. It is thus a measure of the reflecting properties of the target. The value of the RCS of a target is highly dependent on its geometry and materials but also on the frequency band and polarization used for the radar observation. Equation (2.1) shows that compared to other sensing schemes, the power received from targets measured by a radar can vary greatly, depending on the material of the target or its location.

Since WW2, where it was only restricted to the military, the use of radar has reached multiple fields and applications. Although discrete, it is now ubiquitous in our daily lives, from the *"beloved"* Doppler radars used by the police here on Earth, to the Synthetic Aperture radar in space that are used

for monitoring ships [Bru+11], the environment [ZLT12] and illegal deforestation [Wat+18]. More recently, thanks to advances in radar technology that allow for higher carrier frequency and bandwidth in the TeraHertz band [Sta+16], radar are now used in cars, helping to reach autonomous driving [Bil+19]. The higher frequency also generates technologies that are smaller and smaller, for example, *Google* has now multiple products with radars embedded in them [Wan+16] such as smartphones and home appliances. The following sections present a short overview of the different modulations used in radar.

2.2 Basic Pulse Radar

We now review the basic principle of the pulse radar. Let us consider a scene, see Fig. 2.1, where a static target is located at a range R. In order to estimate the range, one can emit a pulse of duration T_p, and then record the echo coming from the environment. The transmitting signal can be

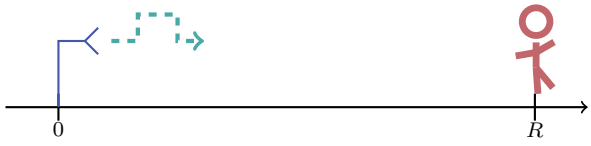


Figure 2.1: A single target scene at a range R from the pulse radar.

expressed as

$$s(t) = \sqrt{P}\Gamma_{T_p}(t),$$

with P the transmitting power that also encapsulates the different gains in the radio frequency chain, and $\Gamma_T(t)$ the window function of length T. For the sake of simplicity, we consider that the antennas have a flat frequency response. The consequence of considering non-ideal antennas is explained at the end of this chapter.

The received signal can be expressed as :

$$r(t) = s(t-\tau) = s(t - \frac{2R}{\mathsf{c}}),$$

with c being the speed of light. One can directly observe that the accuracy of the estimated delay depends directly on the sampling frequency $\frac{1}{T_s}$ used to acquire the received signal.

$$\delta_R = \frac{\mathsf{c}T_s}{2},$$

where δ_R is the range accuracy.

From this simple equation, one can observe that in order to reach a high accuracy using a pulse, one should use extremely high sampling frequency. For example, to have an accuracy of $\delta_R \sim 1\text{m}$, the sampling frequency should be of the order of 100MHz. One can also notice that this sampling frequency does not depend at all on the desired maximum range. The width of the pulse dictates the resolution with which one can estimate targets. Using a matched filter on the received signal to estimate the positions of targets, *i.e.*, $|r(t) * s(t)|$, the width of the resulting ambiguity function is a multiple of T_p and its shape is represented in Fig. 2.2.

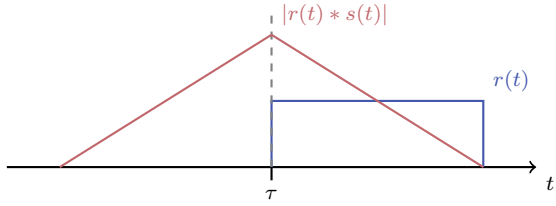


Figure 2.2: Graphical representation of the received signal $r(t)$ and its associated matched filter with the transmitted signal $s(t)$.

In order to increase the accuracy of the estimation using this modulation, one needs to shorten the pulse $\Gamma_{T_p}(t)$, *i.e.*, reduce T_p. This basic system, while being fairly simple and omitting some other hardware considerations, allows us to highlight two issues when using such pulse radars: (*i*) the total energy transmitted PT_p will dictate the SNR that can be attained. At constant SNR, shortening the pulse thus increases the amplitude of the pulse, which requires more complex hardware able to generate high power pulses of short duration. (*ii*) Recovering the delay $\tau = \frac{2R}{\mathsf{c}}$ directly from the measurements requires extremely high sampling rate if sub-meter accuracy is desired. These two effects make the use of pulse radar an expensive endeavour that is not often suited for low power and low cost applications.

This drove radar research to develop other modulations, one of them being the Frequency-Modulated-Continuous-Wave Radar that is the focus of the next section.

2.3 FMCW Radar Signal Model

In this section, we focus on Frequency Modulated Continuous Wave radars with one transmitting and one receiving antenna, this model is extended to 2-D localization in Chapter 5. The radar's transmitting antenna emits a signal $s(t)$ modelled as

$$s^{RF}(t) = \sqrt{P_t} \exp\Big(\mathrm{i}\, 2\pi \big(\int_0^t f_c(\xi) \mathrm{d}\xi \big) + \mathrm{i}\, \phi_0 \Big) \Gamma_{T_c}(t \bmod T_p), \tag{2.2}$$

where P_t is the transmitted power, $f_c(t)$ the transmitted frequency pattern, and $\phi_0 \in [0, 2\pi]$ is the initial phase of the oscillator. The chirp has a duration T_c and is repeated at a *Pulse Repetition Frequency* (PRF) of $\frac{1}{T_p}$. The superscript $(\cdot)^{RF}$ is added to highlight the fact that this signal is centred around a carrier frequency, which, in common radar applications such as automotive, is in the order of GHz (*e.g.*, the *Ka*-band is around 24GHz [Sko80]).

The carrier frequency pattern $f_c(t)$ of an FMCW radar can be characterized as a saw-tooth function (see Fig. 2.3):

$$f_c(t) = f_0 + \frac{B}{T_c}(t \bmod T_p), \tag{2.3}$$

with f_0 the central frequency, and B the spanned bandwidth. Note that, in practical applications, B is not a design parameter but a constraint imposed by government regulations. Considering, for now, a simple ranging problem where the scene is made of only one target located at a range R from the radar which has one receiving antenna, as in Fig. 2.1. The received signal is a time-delayed version of the transmitted signal (2.2),

$$r^{RF}(t) = \alpha s(t - \tau),$$

and is represented in Fig. 2.3 (see the **red** curve). The delay τ is simply the round-trip between the radar and the target and can be thus expressed as

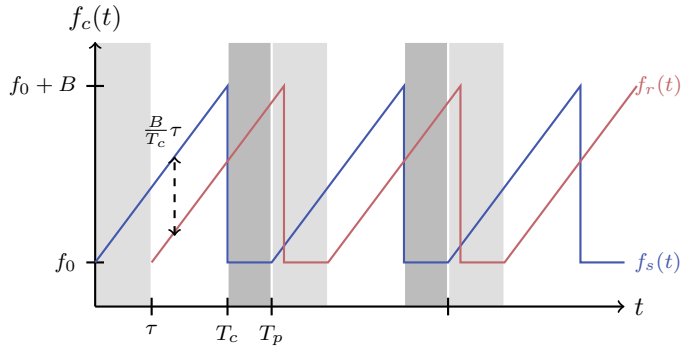


Figure 2.3: Representation of the frequency content of the transmitted (**blue**) and received (**red**) chirps with a delay τ

$\tau = \frac{2R}{c}$. The variable $\alpha \in \mathbb{C}$ represents the amplitude of the received signal, this encompasses the Radar Cross Section (RCS) of the target but also all the other losses (*e.g.*, path losses as in (2.1), and gains in the acquisition chain, ...). For the sake of simplicity, in the rest of our presentation, the complex value α will refer to a global constant amplitude that may change from one line to the other in the description of the reception and demodulation processes. Indeed, each of the these stages is associated with its respective complex gain. It is interesting to note that in Fig. 2.3, the delay τ in the time domain between $s(t)$ and $r(t)$ can also be observed as a frequency shift $\frac{B}{T_c}\tau$ in the frequency domain. From (2.2), the received analytical signal in radio frequency can be expressed as:

$$r^{RF}(t) = \alpha \exp\left(\mathrm{i}\, 2\pi\left(\int_0^{t-\tau} f_c(\xi)\mathrm{d}\xi\right) + \mathrm{i}\,\phi_0\right)\Gamma_{T_c}(t - \tau \bmod T_p). \tag{2.4}$$

The spectrum of (2.4), although narrowband, is centred around f_0 which is, in the case of *Ka-band* radar, in the tens of GHz. Directly sampling this signal is inadvisable as it would require prohibitively high sampling rate that would then incur complex hardware and high costs. One possibility would be to instead sample the signal around its central frequency $f_0 + \frac{B}{2}$ by multiplying the received signal in (2.4) by the carrier signal $\exp\left(-\mathrm{i}\, 2\pi(f_0 + \frac{B}{2})t\right)$. As depicted in Fig. 2.4 the sampling requirements have lowered from the order of the carrier frequency (in GHz) to the order of the bandwidth B of

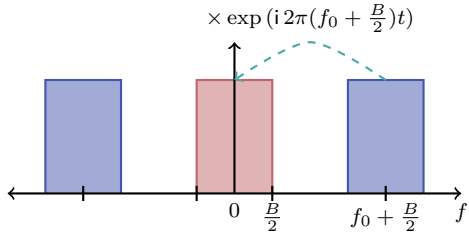


Figure 2.4: Spectral representation of $r(t)$ and $r(t)\exp(-\mathrm{i}(f_0 + \frac{B}{2})t)$.

the modulated signal, *i.e.*, in the order of the hundreds of MHz.

In essence, using this demodulation process, we are now only recording the changes that are imparted by the environment. However, sampling at multiple hundreds of MHz might still be prohibitively high compared to the observation made earlier in Fig. 2.3 that, for small ranges, the delay of interest manifests itself as a frequency shift between the received and transmitted signals. In order to leverage that fact, FMCW radars use coherent demodulation, where to recover this frequency shift, the received signal $r^{RF}(t)$ is demodulated using the transmitted signal $s(t)$ itself. The architecture required to perform this coherent demodulation is presented in Fig. 2.5, where the I and Q channels represent the real and imaginary part of the demodulated received signal $r(t)$ respectively. The process represented in

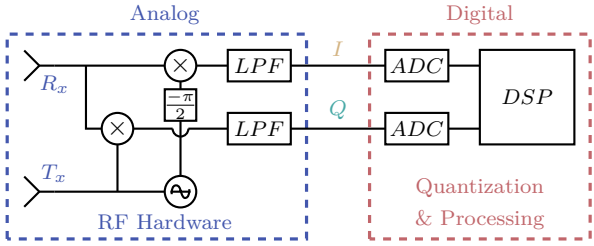


Figure 2.5: Representation of a IQ coherent demodulation for an FMCW radar with one transmit and one receive antenna

Fig. 2.5, can be expressed mathematically simply as

$$r(t) = r^{RF}(t)s^*(t). \tag{2.5}$$

Using the expression of the transmitted signal (2.2), (2.5) reduces to

$$r(t) = \alpha \exp\big(-\mathrm{i}\, 2\pi \int_{t-\tau_0}^{t} f_c(\xi)\mathrm{d}\xi\big)\Gamma_{T_c}\big((t-\tau) \bmod T_p\big)\Gamma_{T_c}(t \bmod T_p). \quad (2.6)$$

As mentioned earlier, assuming a delay $\tau \ll T_c$ small enough such that $\Gamma_{T_c}(t-\tau \bmod T_p) \approx \Gamma_{T_c}(t \bmod T_p)$ and using the saw-tooth model (2.3), the integral in (2.6) becomes

$$\int_{t-\tau_0}^{t} f_c(\xi)\mathrm{d}\xi = \int_{t-\tau_0}^{t} (f_0 + \frac{B}{T}\xi)\mathrm{d}\xi = \tau_0 f_c(t) - \frac{B}{2T}\tau_0^2. \quad (2.7)$$

Combining (2.6) with (2.7) allows us to express the received signal $r(t)$ in base-band, *i.e.*, we have

$$r(t) = \alpha \exp\big(-\mathrm{i}\, 2\pi\tau f_c(t)\big)\Gamma_{T_c}(t \bmod T_p), \quad (2.8)$$

where α also encompasses the static phase-shift $-\frac{B}{2T}\tau_0^2$ in (2.7). In words, (2.8) shows that the coherent demodulation expresses the time difference τ coming from the target as a frequency shift between the transmitted and received signals. This frequency shift linked to the range is represented in Fig. 2.3.

It is interesting to note that this frequency shift $\frac{B}{T_c}\tau$ is often several orders of magnitude below the bandwidth B. This fact motivates the use of the coherent demodulation as the sampling frequency is now linked to $\frac{B}{T_c}\tau$ instead of B. For example, in the context of K-band radar with $f_0 = 24$Ghz and $B = 250$MHz, a target at range of $300m$ will generate a frequency shift of 200kHz for a PRF of 12kHz. The resolution with which a target can be estimated is not linked to the duration of the transmitted pulse, *i.e.*, T_c or the PRF but is related to the bandwidth B of the chirp signal $f(t)$. The resolution δ_R can be simply expressed as $\delta_R = \frac{\mathsf{c}}{2B}$ thanks to the property of the Fourier transform. We illustrate this by giving an example in Fig. 2.6. Two Fourier transforms of $r(t)$ for a target located at $20m$ with two different bandwidths $B = 30$Mhz in **blue** and $B = 12$MHz in **green** are represented. For *Ka-band* radar, the regulations allows for a bandwidth of 250MHz which gives a resolution of $0.6m$.

Considering a simple additive model, the received base-band signal from

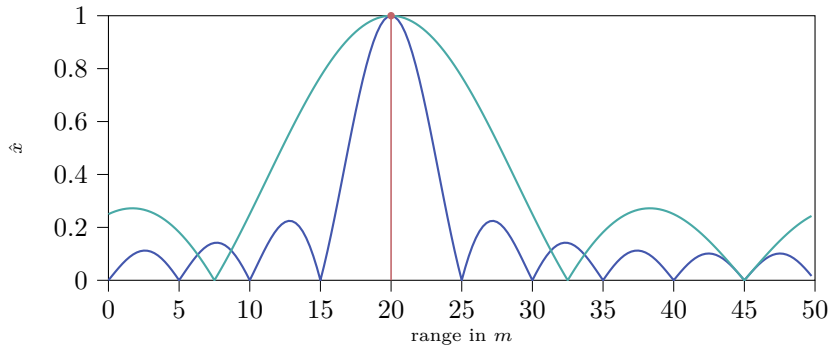


Figure 2.6: Representation of the Fourier transform of a baseband received signal with different Bandwidths (30MHz and 12MHz) for a target located at a $20m$ range

a scene with s targets can be expressed as

$$r(t) = \sum_{i}^{s} \alpha_i \exp\big(-\mathrm{i}\, 2\pi \frac{2R_i}{\mathsf{c}} f_c(t)\big)\Gamma_{T_c}(t \bmod T_p), \tag{2.9}$$

where α_i and R_i are the received powers and ranges from each of the s targets measured in the scene. If the duration of the acquisition is restricted to $t \in [0, T_c]$, one can see that the problem of estimating the ranges of these s targets is tantamount to estimating the frequency content of $r(t)$ and by associating each peak in the frequency domain to a specific range. Chapter 4 will describe in detail the acquisition process and its effect on the accuracy and precision on the range estimation problem.

This model is remarkably simple in its interpretation of the estimation process. It is however important to acknowledge all of the non-idealities that are not accounted within (2.9). It is assumed that the frequency response of the antennas and of the local oscillator that generates the chirp are flat for the considered bandwidth B. In practice, the frequency response of these components are never exactly flat and this impacts the quality of the baseband signals [SSS06; Vin+11]. Indeed, the received signal in (2.8) is affected in the following fashion

$$r(t) = \alpha h(f_c(t)) \exp\big(-\mathrm{i}\, 2\pi\tau f_c(t)\big)\Gamma_{T_c}(t \bmod T_p),$$

where $h(f)$ is the combined frequency response of the system that encom-

passes the transmitting and receiving antennas as well as the RF hardware. Given the fact that the estimation process, in its simplest sense, is a Fourier transform, the perfect target response corresponding to a delta centred on the range R will be tainted by $H(d) = \int_{f_0}^{f_0+B} h(\xi) \exp\left(-\mathrm{i}\, 2\pi\xi \frac{2d}{\mathrm{c}}\right) d\xi$. In extreme cases, this could impair the proper estimation of the s different targets as these responses $H(d)$ might overlap.

The model in (2.9) also assumes point-like targets. While this assumption might be valid in cases where resolution δ_R is coarser than the dimension of the targets itself. In most applications, however, resolutions that are thinner then the target are often desired. The effect of the noise and the clutter generated by the environment has also been omitted in this model, Chapters 5 and 7 will highlight its effect on the particular case of the 1-bit quantization of these noisy measurements.

Finally, the coherent demodulation process is assumed to be perfect and to have a perfect isolation between the transmit and receive circuits. In practice as depicted in Fig. 2.7, leakages between these two circuits might happen. This, in turns, affects the coherent demodulation as the RF received signal is now a weighted sum between the actual echo from the target and the time delayed leakage signal [Haf+20]:

$$r^{RF}(t) = \alpha s(t-\tau) + \sum_i \beta_i s(t - \tau_{i,\text{leakage}}),$$

where β_i is the amplitude, and $\tau_{i,\text{leakage}}$ of the different leakages that can occur in the radar system.

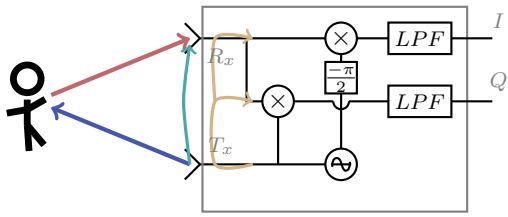


Figure 2.7: Schematic representation of leakage between the transmit and received circuit of the radar. Direct coupling between the antennas is represented in **green** and within the circuit is represented in **yellow**.

After the coherent demodulation, the baseband signal can now be de-

noted as

$$r(t) = \alpha \exp\left(-\mathrm{i}\, 2\pi\tau f_c(t)\right) + \sum_i \beta_i v_i(t),$$

where $v_i(t) = \exp\left(-\mathrm{i}\, 2\pi\tau_{i,\text{leakage}} f_c(t)\right)$ is the low frequency signal of one source of leakage i. As $\tau_{i,\text{leakage}} \ll \tau$, this low frequency signal thus disrupts the received signal and, in some cases, might have a dynamic larger than the signal of interest which also hinders its acquisition. Indeed, adjusting the dynamic range of the ADCs to this larger dynamic reduces the effective number of bits devoted to the signals coming from the targets. This leakage can often be corrected during the processing in classical acquisition scheme.

Chapter 3

Compressive Sensing

The work presented in this thesis relies heavily on the framework of Compressive Sensing and its quantized counterpart Quantized Compressive Sensing. Consequently this chapter constitutes a non-exhaustive summary of this field with a particular attention to the tools, notions, and theorems used throughout this thesis. The problem of estimating sparse vectors from the compressed measurements is first introduced in the context of Fourier transform, before being generalized to linear systems. To that end, the notion of sparse vectors is briefly reviewed and the requirements on the measurement process is explicited using the Restricted Isometry Property. Different flavours of reconstruction algorithms are then presented before concluding with the introduction of the different quantizers commonly used in QCS.

3.1 Problem Statement

As seen in the previous chapter regarding radar signal models, estimating the range of targets using an FMCW radar is equivalent to estimating the frequency content of the received signal (see (2.9)). These analogic signals must be first measured and sampled before one can study their frequency content. One of the most fundamental results of signal processing is Shannon's sampling theorem [Sha49] that says:

Theorem 3.1 (Shannon's sampling theorem, Theorem 1 in [Sha49])**.** *If a function $s(t)$ contains no frequencies higher than B Hertz, it is completely determined by giving higher than its ordinates at a series of points spaced $\frac{1}{2B}$ seconds apart.*

Considering a signal $s(t)$ of duration T, the sampling rate defined in Theorem 3.1 will generate $2TB$ samples, *i.e.*, $s[m] = s(m\frac{1}{2B})$, with $m \in [2TB]$. With $s[m]$, one can now estimate its spectral content $S[n]$ using the Discrete Fourier Transform (DFT):

$$S[n] = \sum_m s[m] \exp\left(-\mathrm{i}\, 2\pi \frac{mn}{N}\right),$$

with $N = 2TB$. Furthermore, its continuous representation in the frequency domain can be computed using a simple *sinc* interpolation at each of the points in $S[n]$.

This is the foundation on which signal processing is built on. But this requirement does not depend on the actual frequency content of $S(f)$. Indeed, only the bandwidth occupied by the signal defines the sampling frequency, not the structure of the signal. In this seminal work that triggered an enthusiasm for the study of compressive sensing, authors in [CRT06a], showed that in the context of band limited signals (in the complex domain) that are composed of only $s \ll N$ frequency tones, *i.e.*,

$$s(t) = \sum_i^s x_i \exp(\mathrm{i}\, 2\pi f_i t),$$

with $x_i \in \mathbb{C}$ and $f_i \in [-\frac{B}{2}, \frac{B}{2}]$ being the amplitude and frequency associated to the $i^{\text{th}} \in [s]$ frequency tone, it is possible to reconstruct the spectral information $S[n]$, from incomplete time domain measurements. In that case, they showed that the number of measurements m needed for the reconstruction of $S[n]$ using convex optimisation can be lowered to $m \leq Cs \log(N) \leq N$, with $C > 0$. In other words, if the signal has s non-zero components in the frequency domain, one can relate the minimum number of measurements m to s instead of N, *i.e.*, by sub-sampling the signal in the time domain. As it

will be introduced in the next sections, Compressive Sensing is the generalisation of this fact to a broader class of problems where the measurements can be represented in a low-dimensional space.

3.2 Compressive Sensing

The section above hinted that the number of measurements prescribed by Shannon's theorem, in Theorem 3.1, could be lowered provided some assumptions on the signal of interest (*e.g.*, few non-zero components in the frequency domain) thanks to the theory of Compressive Sensing. The rest of this chapter is intended to be an introduction to this field.

We start by considering the following linear model:

$$\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x}, \tag{3.1}$$

with $\boldsymbol{y} \in \mathbb{C}^m$ representing the measurements, $\boldsymbol{x} \in \mathbb{C}^N$ is the signal that one wants to estimate from the measurements, and the matrix $\boldsymbol{\Phi} \in \mathbb{C}^{m \times N}$ that represents the linear measurement process of the signal $\boldsymbol{x}$, with $m < N$.

One could try to directly solve the linear model defined in (3.1) but because $m \leq N$, the system is undetermined. This means that without any other assumptions on the model, there is potentially an infinite number of solutions to this problem. Compressive Sensing attempts to solve this issue.

The spirit of Compressive Sensing and the different ingredients that are usually involved, can be expressed in the following recovery algorithm.

Theorem 3.2 (Basis Pursuit Denoising, from Theorem 6.12 in [FR13])**.** *If a matrix $\boldsymbol{\Phi} \in \mathbb{C}^{m \times N}$ follows the (ℓ_2, ℓ_2)-Restricted Isometry Property of order $2s$ with a constant δ such that*

$$\delta < \frac{4}{\sqrt{41}} \approx 0.6246.$$

Then, for any vector $\boldsymbol{x} \in \mathbb{C}^N$ and its associated noisy measurements $\boldsymbol{y}$ with

$\|\boldsymbol{\Phi x} - \boldsymbol{y}\|_2 \leq \eta$, *a solution* $\boldsymbol{x}^{\#}$ *of*

$$\min_{\boldsymbol{z} \in \mathbb{C}^N} \|\boldsymbol{z}\|_1 \ s.t \ \|\boldsymbol{\Phi z} - \boldsymbol{y}\|_2 \leq \eta$$

approximates the vector $\boldsymbol{x}$ *with an error upper-bounded by*

$$\|\boldsymbol{x} - \boldsymbol{x}^{\#}\|_2 \leq \frac{C}{\sqrt{s}} \sigma_s(\boldsymbol{x})_1 + D\eta,$$

where the constants C,$D > 0$ *depend only on* δ.

The result of Theorem 3.2 is striking. In words, this theorem shows that the Basis Pursuit Denoising algorithm is robust to noise and stable when reconstructing vectors that are not exactly sparse (the meaning of $\sigma_s(\cdot)_1$ will be explicited later) in a specific setting. What is more, when the signal of interest is noiseless and sparse, the BPDN algorithm provides perfect reconstructions, and this in the challenging $m \leq N$ setting. Compressive Sensing has been used in numerous applications : radar [End13], MRI [LDP07], and many others (see [FR13; RDD18] and reference therein). Let us now review the different ingredients that are required by BPDN.

3.2.1 Sparsity

The noiseless reconstruction error of BPDN depends on the term $\frac{C}{\sqrt{s}} \sigma_s(\boldsymbol{x})_1$, with

$$\sigma_s(\boldsymbol{x})_1 = \min_{\boldsymbol{z} \in \bar{\Sigma}_s^N} \|\boldsymbol{x} - \boldsymbol{z}\|_1, \tag{3.2}$$

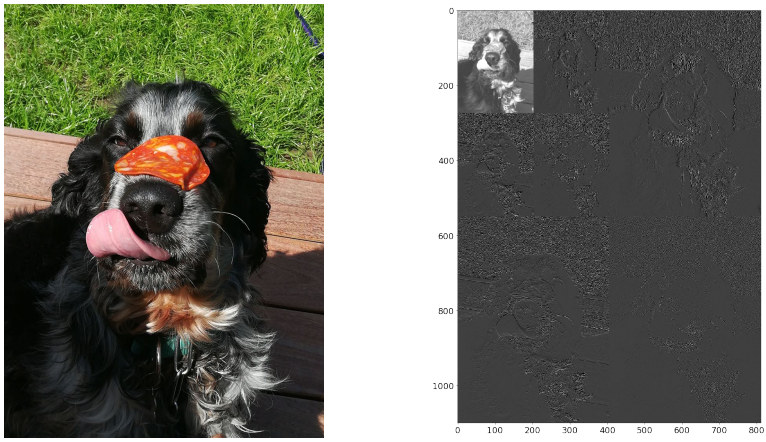
where the set $\bar{\Sigma}_s^N$ is the set of complex s-sparse vectors. $\sigma_s(\boldsymbol{x})_1$ measures how close the vector $\boldsymbol{x}$ is to being s-sparse. For a vector to be s-sparse, it must not have more than s non-zero components, *i.e.*, $|\operatorname{supp}(\boldsymbol{x})| = \|\boldsymbol{x}\|_0 \leq s$. Fig. 1.2 provides an example of a signal that is approximately sparse. Indeed, the *Range-Doppler* map is made of a few dominant components corresponding to the different cars measured by the radar. These approximately sparse signals, however, are not what is directly measured by the radar. Indeed, the previous chapter showed that using an FMCW radar, the measured signal is the sum of different complex exponentials (see (2.9)).

The supports of these received signals are far from sparse, only their spectrum are. In fact, in numerous applications, the signals of interests are rarely sparse in the domain of acquisition.

Like radar signals, images acquired by a camera sensor are not sparse, but can have a representation that is sparse in another domain. The signal $\boldsymbol{x}$ can be represented as

$$\boldsymbol{x} = \boldsymbol{\Psi}\boldsymbol{\alpha},$$

where $\boldsymbol{x}$ is the signal of interest, $\boldsymbol{\alpha}$ is the sparse representation of $\boldsymbol{x}$ using the appropriate representation $\boldsymbol{\Psi}$. For example, using the wavelet transform, the Fig. 3.1a can be represented as Fig. 3.1b.



(a) (b)

Figure 3.1: (a) Picture of a *not* so well-behaved dog. (b) Wavelet transform of the grayscale version of Fig. 3.1a using Daubechies 1 wavelets with 2 levels.

If one studies the amplitude of the different coefficients of Fig. 3.1b by sorting them according to their amplitude in Fig. 3.2, one can see that most of the energy of the image $\boldsymbol{\alpha}$ are encapsulated in only a few non-zero elements in $\boldsymbol{x}$. The curve in **blue** represents the evolution of the amplitude of the sorted coefficients, while the **red** curve corresponds to the cumulative sum of the coefficient α_i. In fact, this cumulative sum can be linked to the ℓ_1-approximation error in (3.2) as $\sum_{j=0}^{i} |\alpha_j|/\|\boldsymbol{\alpha}\|_1 = 1 - \sigma_i(\boldsymbol{\alpha})_1/\|\boldsymbol{\alpha}\|_1$. Thanks to the steep decrease of the amplitude of these coefficients, one can

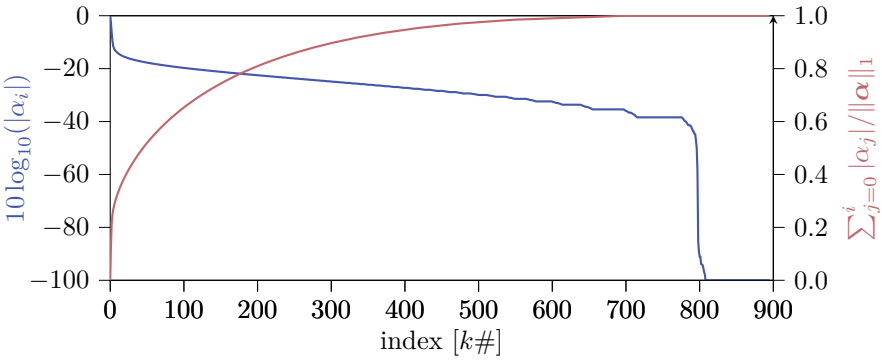


Figure 3.2: Sorted amplitudes of the wavelet coefficient of Fig. 3.1a.

see that more than 70% of the cumulative sum is enclosed in a ninth of the coefficients.

In fact, if one only keeps the biggest coefficients of the wavelet transform and then applies the inverse transform, one can see that most of the image's information is indeed enclosed in these few coefficients. Fig. 3.3 shows the reconstruction of Fig. 3.1a by Hard-Thresholding the coefficients in Fig. 3.2 and applying the inverse transform. In Fig. 3.3a, only 10% of the coefficients are kept and no discernible degradation can be observed. As less and less coefficients remain for the reconstruction, the image's quality deteriorates but the fact that the compression operated in Fig. 3.3d is of 0.1% is impressive. This example, although simple, clearly shows that using the proper representation can significantly help sparsifying signals, provided this representation does exist.

3.2.2 The Restricted Isometry Property

The second condition imposed on the linear system introduced in (3.1) for the reconstruction to be successful is that the measurement matrix $\boldsymbol{\Phi} \in \mathbb{C}^{m \times N}$ that links the measurement $\boldsymbol{y} \in \mathbb{C}^m$ and the sparse vector $\boldsymbol{x} \in \mathbb{C}^N$ must follow the Restricted Isometry Property (RIP). The RIP can be defined as:

(a) 10%

(b) 5%

(c) 1%

(d) 0.1%

Figure 3.3: Inverse wavelet transform (Daubechie 1 at level 4) of Fig. 3.1a where only $10\%, 5\%, 1\%, 0.1\%$ of the coefficient are kept.

Definition 3.1. Given $\delta > 0$, the matrix $\boldsymbol{\Phi} \in \mathbb{C}^{m \times N}$ satisfies the (ℓ_2, ℓ_2)-RIP(s, δ) if, for all $\boldsymbol{x} \in \bar{\Sigma}_s^N$,

$$(1 - \delta)\|\boldsymbol{x}\|_2^2 \leq \tfrac{1}{m}\|\boldsymbol{\Phi}\boldsymbol{x}\|_2^2 \leq (1 + \delta)\|\boldsymbol{x}\|_2^2.$$

The definition in Def. 3.1 is tantamount to saying that the columns of $\boldsymbol{\Phi}$ should be close to orthogonal. The RIP can also be defined as

Remark 3.1. Given $\delta > 0$, the matrix $\boldsymbol{\Phi} \in \mathbb{C}^{m \times N}$ satisfies the (ℓ_2, ℓ_2)-RIP(s, δ) if,

$$\delta = \max_{\mathcal{S} \in [N], |\mathcal{S}| \leq s} \| \boldsymbol{I}_d - \tfrac{1}{m} \boldsymbol{\Phi}_{\mathcal{S}}^H \boldsymbol{\Phi}_{\mathcal{S}} \|$$

This definition can be linked back to the problem of estimating the frequency content of a measured signal. One ubiquitous property of the Fourier transform is what is referred as Parseval's Theorem,

Definition 3.2 (Parseval's Theorem)**.** For a signal $x(t)$ and its associated Fourier representation $X(f)$, the energy of these two signals are linked by

$$\int_{-\infty}^{\infty} |x(t)|^2 dt = \int_{-\infty}^{\infty} |X(f)|^2 df.$$

In words, the Fourier transform of a signal has the same energy as the signal itself and there is an isometry between the time domain and the frequency domain. Discretizing the representation by sampling the signal in the time domain gives us the corresponding expression:

Remark 3.2 (Parseval's Theorem for discrete signals)**.** For a discrete signal $\boldsymbol{x}$ and its associated Fourier representation $\boldsymbol{F}\boldsymbol{x}$, the energy of these two signals are linked by

$$\|\boldsymbol{x}\|_2^2 = \tfrac{1}{m} \|\boldsymbol{F}\boldsymbol{x}\|_2^2.$$

It is important to note that here $\boldsymbol{x} \in \mathbb{C}^N$ and $\boldsymbol{F} \in \mathbb{C}^{N \times N}$, *i.e.*, our linear system is not sub nor over-sampled. Comparing Remark 3.2 to the definition

of the RIP found in Def. 3.1, one could interpret the RIP as an extension of Parseval's theorem in the setting where the number of measurements does not match the dimension of the signal of interest, *i.e.*, $m \neq N$. This generalization, however, comes at a cost in the restriction to the set of s-sparse vector $\bar{\Sigma}_s^N$ for the RIP while Parseval's theorem applies to all signals $\mathbb{C}^N$.

Indeed, for non-sub-sampled Fourier transform Remark 3.2 also means that all the columns of $\boldsymbol{F}$ are orthogonal with each other, *i.e.*,

$$\tfrac{1}{m}\boldsymbol{F}^H\boldsymbol{F} = \boldsymbol{I}_N \rightarrow \forall i \neq j \in [N],\, \langle \boldsymbol{F}_{:,i}, \boldsymbol{F}_{:,j} \rangle = 0$$

this relationship cannot be assumed to hold when the Fourier transform is sub-sampled, $\boldsymbol{\Phi} = \boldsymbol{F}_\Omega$, with $|\Omega| = m \leq N$. In that case

$$\mu = \tfrac{1}{m} \max_{i \neq j \in [N]} |\langle \boldsymbol{F}_{\Omega,i}, \boldsymbol{F}_{\Omega,j} \rangle|$$

which can be far from zero for some $m \ll N$. This measure of the orthogonality of the columns of the sub-sampled matrix is called the coherence. Fig. 3.4 shows the back-projection different measurements $\boldsymbol{F}_\Omega \boldsymbol{x}$, with $\boldsymbol{x}$ be-

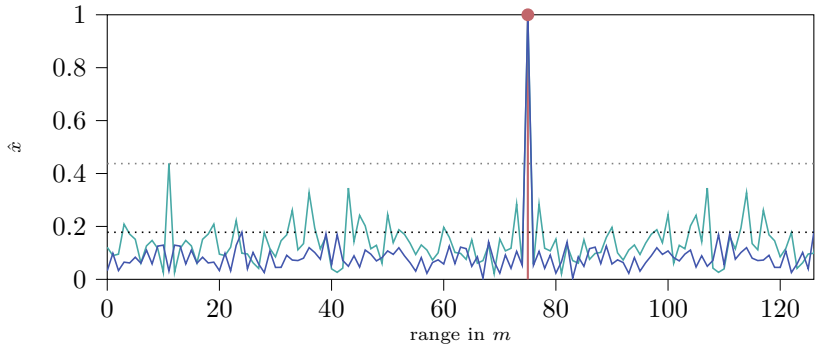


Figure 3.4: Back-projection of Partial Fourier measurement $\frac{1}{m}|\boldsymbol{\Phi}^H\boldsymbol{\Phi}\boldsymbol{x}|$; the original 1-sparse $\boldsymbol{x} \in \mathbb{C}^{128}$ is located in 75 and the signals in **blue** and **green** are the reconstruction for $m = \frac{N}{2} = 64$ and $m = \frac{N}{4} = 32$ respectively; the dotted line represent the coherence μ obtained for each sub-sampling.

ing a 1-sparse vector different amount of sub-sampling $|\Omega| = m \leq N$. As the number of measurements is lowered the mutual coherence of $\boldsymbol{\Phi} = \boldsymbol{F}_\Omega$ increases.

The RIP constant δ can be bounded using the coherence by $\mu \leq \delta \leq (s-1)\mu$ for $s \geq 2$. The coherence of a matrix is lower-bounded by the Welsh bound that states as

Theorem 3.3 (Welsh Bound from Thm.5.7 in [FR13])**.** *The coherence of a matrix $\boldsymbol{\Phi} \in \mathbb{K}^{m \times N}$ with ℓ_2-normalized columns satisfies*

$$\mu \geq \sqrt{\frac{N-m}{m(N-1)}}.$$

Theorem 3.3 tells us that regardless of the sub-sampling or of the matrix used, the coherence cannot decrease faster than $\mathcal{O}(m^{-\frac{1}{2}})$. Conversely, the RIP can be used to upper-bound the coherence. For $s = 2$ the expression of the RIP in Rem. 3.1 corresponds to the coherence.

Estimating the RIP constant δ of deterministic matrices is computationally expensive [FR13]. Compressive Sensing solves this issue by studying instead the properties of random matrices. The RIP property has been established for Gaussian and sub-Gaussian matrices and more importantly in the setting of radar signal processing, for randomly sub-sampled Fourier transform.

3.2.3 Reconstruction Algorithms

Given that the aim of Compressive Sensing is to estimate a sparse vector $\boldsymbol{x} \in \bar{\Sigma}_s^N$ from measurements $\boldsymbol{y} = \boldsymbol{\Phi x}$, a natural reconstruction strategy would be

$$\min_{\boldsymbol{u} \in \mathbb{C}^N} \|\boldsymbol{u}\|_0 \text{ such that } \|\boldsymbol{y} - \boldsymbol{\Phi u}\|_2 \leq \eta$$

Solving this, however, is not computationally tractable [FR13; Don06b]. A relaxed version of this problem is to substitute the ℓ_0-norm with the ℓ_1-norm, which corresponds to the Basis Pursuit Denoising introduced in Theorem 3.2.

There exist other methods that can solve the problem of recovering a sparse vector from measurements. Projected-Back-Projection (PBP) is one of the simplest algorithms and is defined as

Definition 3.3 (Projected-Back-Projection)**.** For a measurement vector $\boldsymbol{y} \in \mathbb{C}^m$ and given a linear model with a measurement matrix $\boldsymbol{\Phi} \in \mathbb{C}^{m \times N}$, one can estimate a s-sparse vector using the following algorithm,

$$\hat{x} = \tfrac{1}{m} \mathsf{H}_s\big(\boldsymbol{\Phi}^H \boldsymbol{y}\big),$$

where $\mathsf{H}_s(\cdot)$ is the hard-thresholding operator that keeps the s biggest elements in amplitude.

Which is simply applying a matched filter on measurement $\boldsymbol{y}$ and keeping the s-biggest elements (in amplitude). Although extremely simple, this algorithm has a uniform bound on the ℓ_2-reconstruction for all s-sparse vectors.

Theorem 3.4 (PBP reconstruction [FR13])**.** *Using the PBP algorithm defined in Def. 3.3 and its estimate $\hat{\boldsymbol{x}}$, one can upper-bound the ℓ_2 reconstruction error for all vectors $\boldsymbol{x} \in \tilde{\Sigma}_s^N$ by*

$$\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2 \leq 2\delta,$$

if the measurement matrix follows the (ℓ_2, ℓ_2)-RIP$(2s, \delta)$.

Proof. Starting with the definition of the PBP estimate, with $\mathcal{S} = \operatorname{supp}(\hat{\boldsymbol{x}})$, one can upper-bound the reconstruction by

$$\begin{aligned} \|\boldsymbol{x} - \tfrac{1}{m} \mathsf{H}_s\big(\boldsymbol{\Phi}^H \boldsymbol{y}\big)\|_2 &= \|\boldsymbol{x} - \tfrac{1}{m} \big(\boldsymbol{\Phi}^H \boldsymbol{y}\big)_{\mathcal{S}}\|_2 \\ &\leq \|\boldsymbol{x} - \tfrac{1}{m} \big(\boldsymbol{\Phi}^H \boldsymbol{y}\big)_{\mathcal{T}}\|_2 + \tfrac{1}{m} \|\big(\boldsymbol{\Phi}^H \boldsymbol{y}\big)_{\mathcal{S}} - \big(\boldsymbol{\Phi}^H \boldsymbol{y}\big)_{\mathcal{T}}\|_2, \end{aligned} \tag{3.3}$$

with $\mathcal{T} = \operatorname{supp}(\boldsymbol{x}) \cup \mathcal{S}$, with $|\mathcal{T}| \leq 2s$. Because $\big(\boldsymbol{\Phi}^H \boldsymbol{y}\big)_{\mathcal{S}}$ is the best s-sparse approximation of $\boldsymbol{\Phi}^H \boldsymbol{y}$, one can bound the second term of (3.3) by the first. The bound becomes

$$\|\boldsymbol{x} - \tfrac{1}{m} \mathsf{H}_s\big(\boldsymbol{\Phi}^H \boldsymbol{y}\big)\|_2 \leq 2\|\boldsymbol{x} - \tfrac{1}{m} \big(\boldsymbol{\Phi}^H \boldsymbol{y}\big)_{\mathcal{T}}\|_2.$$

Using the definition of RIP in Remark 3.1 and the fact that $\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x}$, one can finally bound

$$\begin{aligned}\|\boldsymbol{x} - \tfrac{1}{m}\mathsf{H}_s\big(\boldsymbol{\Phi}^H\boldsymbol{y}\big)\|_2 &= 2\|\big(\boldsymbol{x} - \tfrac{1}{m}\boldsymbol{\Phi}^H\boldsymbol{\Phi}\boldsymbol{x}\big)_{\mathcal{T}}\|_2 \\ &\leq 2\|\big(\boldsymbol{I}_d - \tfrac{1}{m}\boldsymbol{\Phi}^H\boldsymbol{\Phi}\big)_{\mathcal{T}}\|\|\boldsymbol{x}\|_2 \leq 2\delta.\end{aligned}$$

□

The reconstruction provided by PBP is directly linked to the RIP constant of $\boldsymbol{\Phi}$. This proof was also included to showcase the basic tools that are used in the different chapters.

One can build upon this first estimate by trying to enforce a consistency between the current estimate and the measurements used for the reconstruction.

Definition 3.4 (Iterative-Hard-Thresholding, [BD09])**.** The k^{th} iteration of the IHT algorithm is expressed as

$$\hat{\boldsymbol{x}}^k = \mathsf{H}_s\big(\hat{\boldsymbol{x}}^{k-1} + \tfrac{1}{m}\boldsymbol{\Phi}^H(\boldsymbol{y} - \boldsymbol{\Phi}\hat{\boldsymbol{x}}^{k-1})\big),$$

where $\hat{\boldsymbol{x}}^0$ is the PBP estimate.

Leveraging again the RIP property of $\boldsymbol{\Phi}$, one can also upper-bound the reconstruction of IHT.

Theorem 3.5 (Reconstruction guarantee of IHT, Corr.1 from [BD09])**.** *Given a noisy observation $\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x} + \boldsymbol{e}$, where $\boldsymbol{x} \in \mathbb{C}^N$ and $\tilde{\boldsymbol{x}} \in \bar{\Sigma}_s^N$ being the best s-sparse approximation of $\boldsymbol{x}$. If $\boldsymbol{\Phi}$ has the (ℓ_2, ℓ_2)-RIP$(3s, \delta)$ with $\delta < \frac{1}{8}$, then, at iteration k, IHT will recover an approximation $\hat{\boldsymbol{x}}^k$ satisfying*

$$\|\boldsymbol{x} - \hat{\boldsymbol{x}}^k\|_2 \leq 2^{-k}\|\tilde{\boldsymbol{x}}\|_2 + 5\tilde{\epsilon}_s,$$

where

$$\tilde{\epsilon}_s = \|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|_2 + \frac{1}{\sqrt{s}}\sigma_s(\boldsymbol{x})_1 + \|\boldsymbol{e}\|_2.$$

Furthermore, after at most

$$k^\star = \lceil \log_2 \left(\frac{\|\boldsymbol{x}\|_2}{\tilde{\epsilon}_s} \right) \rceil$$

iterations, IHT estimates $\boldsymbol{x}$ *with accuracy*

$$\|\boldsymbol{x} - \hat{\boldsymbol{x}}^{k^\star}\|_2 \leq 6\tilde{\epsilon}_s.$$

What is striking about this result is that thanks to the RIP property of the measurement matrix, one can reconstruct sparse signals perfectly while being in the challenging setting of $m \leq N$.

3.3 Quantized Compressive Sensing

In this section, a brief overview of the quantization process is presented. The motivation of including this process in the signal model is explicited before reviewing the most common type of quantizers and their properties. The chapter finishes with the introduction of the dithering, which is one of the central focus of this thesis.

Up to now, the measurements $\boldsymbol{y}$ used for the reconstructions were treated as variables existing in a continuous space such as $\mathbb{C}^m$. However, before any signals can be processed, it needs to be digitized. This is done using ADCs (Analogic to Digital Converters) that convert continuous signals coming from the sensor in volts to a digital representation expressed in bits. This is this digital representation of the analogic signal that is used in the reconstruction process. These ADCs have a finite resolution represented by the number of bits b on which the acquired measurements are represented as well as a dynamic range beyond which the digital measurements saturate. This induces a discrepancy between the continuous signal and its quantized counter-part, even in a noiseless setting.

The linear model introduced in (3.1) becomes

$$\boldsymbol{z} = \mathcal{A}(\boldsymbol{\Phi}\boldsymbol{x}),$$

where $\mathcal{A}(\cdot)$ is the quantizer. For the purpose of this introduction, we restrict our presentation to uniform scalar quantizers. There exist however other quantizers such as the $\Sigma\Delta$ quantization that has been studied in QCS in [Bou+15; Gun+10] (and see reference contained therein). The most common quantizer is the mid-rise quantizer defined as

$$\mathcal{Q}_\epsilon^b(\lambda) = \begin{cases} (2^b - 1)\frac{\epsilon}{2} & \text{if } \lambda \geq (2^b - 1)\frac{\epsilon}{2} \\ \lfloor \frac{\lambda}{\epsilon} \rfloor \epsilon + \frac{\epsilon}{2} & \text{if } |\lambda| < (2^b - 1)\frac{\epsilon}{2} \\ -(2^b - 1)\frac{\epsilon}{2} & \text{if } \lambda < -(2^b - 1)\frac{\epsilon}{2} \end{cases}$$

This function has a resolution of ϵ and the measurements are recorded using b bits. The quantizer has thus a dynamic range of $2^b\epsilon$ centred around 0 before saturating. Depending on the resolution and dynamic of this quantizer with respect to the measurements that need to be acquired, the effect of this quantization can be ignored. One naive way of dealing with the effect

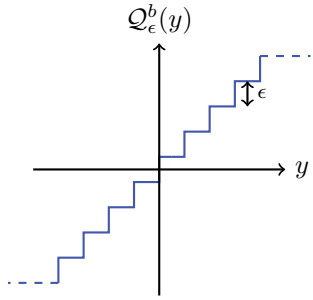


Figure 3.5: Representation of the mid-rise quantizer of resolution ϵ with b bits.

of this quantizer is to consider the discrepancy it induces as a noise, *i.e.*,

$$\boldsymbol{z} = \mathcal{Q}_\epsilon^b(\boldsymbol{\Phi}\boldsymbol{x}) = \boldsymbol{\Phi}\boldsymbol{x} + \boldsymbol{q}. \tag{3.4}$$

If the measured signal does not saturate the quantizer, then this quantization noise is bounded by the resolution of the quantizer, *i.e.*, $\|\boldsymbol{q}\|_\infty \leq \frac{\epsilon}{2}$ which in turns means that $\|\boldsymbol{q}\|_2 \leq \sqrt{m}\frac{\epsilon}{2}$. There is a trade-off, however, for ADCs between the resolution and the sampling frequency [Wal99; Le+05; ZO18]. So, restricting the acquisition of signals to only high-resolution ADCs might result in more power or cost demanding hardware.

Futhermore, all ADCs have a finite dynamic range and thus suffer from saturation. This effect on the reconstruction of sparse signals was studied in [Las+11; FL18], where modifications of the reconstruction algorithms were studied to include explicitely this effect, for example through a consistency condition, or by ignoring the saturated measurements by relying on the *democratic* property of the measurement matrix.

This thesis however, studies quantizers that are extremely coarse where using these assumptions would result in poor reconstruction performances.

3.3.1 One-bit Quantization

We now study the coarsest quantization that exists: the 1-bit quantization. It is defined, for real, signals as

$$\mathcal{Q}_\epsilon(\lambda) = \mathcal{Q}_\epsilon^1(\lambda) = \frac{\epsilon}{2}\operatorname{sign}(\lambda).$$

In this setting, the quantized measurements are now all saturated and the information regarding the norm of λ is completely lost as illustrated in Fig. 3.6a. As seen in Chapter 2, received radar signals are complex thanks to the demodulation process and thus both the real and imaginary part of the signal need to be acquired. The quantization of complex signals is performed as

$$\mathcal{Q}_\epsilon(\lambda^{\mathbb{R}} + \mathrm{i}\,\lambda^{\mathbb{I}}) = \frac{\epsilon}{2}\operatorname{sign}(\lambda^{\mathbb{R}}) + \mathrm{i}\,\frac{\epsilon}{2}\operatorname{sign}(\lambda^{\mathbb{I}}). \tag{3.5}$$

The quantization process in (3.5), represented in Fig. 3.6b, effectively only records in which quadrant in the complex plane the measurements are lying. Given these extremely coarse measurements, using the approximation in (3.4) is tantamount to considering a quantization noise that is as big as the signal itself. Furthermore, this quantization noise is not independent from the signal itself. Although classical Compressive Sensing methods are robust to noise, ignoring the effect of this quantization limits the reconstructions severely.

One-bit CS was first introduced in [BB08] and was then the subject of numerous studies (see [Dir19; Bou+15] and the references therein). This challenging acquisition setting showed that the common tools of CS such as the (ℓ_2, ℓ_2)-RIP were not sufficient to obtain high quality reconstruc-

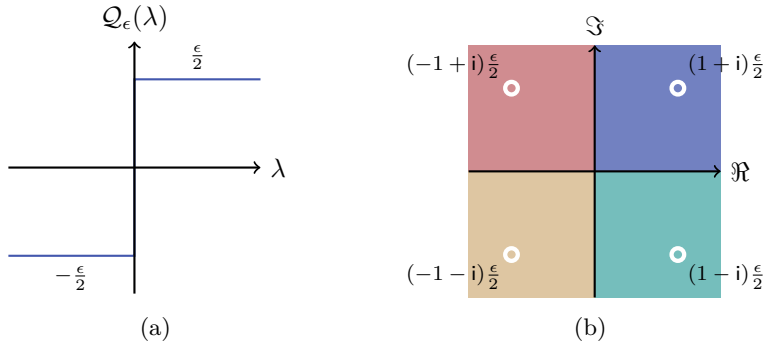


Figure 3.6: (a) Graphical representation of $\mathcal{Q}_\epsilon(\lambda)$. (b) Extension to the complex domain.

tions. Some research relied on specific properties of the considered linear systems, such as the (ℓ_1, ℓ_2)-RIP in [Fou17], the use of circulant matrices in [DJR19], or the study of other properties such as the *binary-ϵ-stable embedding (BϵSE)* [Jac+13]. While these tools provide interesting insights into the 1-bit CS potentialities, their applicability in a radar context are far from straightforward. Indeed, the radar model, as presented in Chapter 2 do not possess these specific properties such as the (ℓ_1, ℓ_2)-RIP (see Def. 3.1).

3.3.2 Additive Dithering

As mentioned in the previous section, one way of dealing with the effect of this 1-bit quantization is to associate it to a quantization noise. In numerous applications, noisy measurements $\boldsymbol{y}$ can be modeled as being tainted by a Gaussian random vector $\boldsymbol{n} \sim \mathcal{N}^m(0, \sigma^2)$. The common practice in signal processing when encountering such noisy measurements is simply to increase the duration of the acquisition (*i.e.*, m) and leveraging the fact that

$$\mathbb{E}\{\boldsymbol{y} + \boldsymbol{n}\} = \boldsymbol{y}.$$

This property, however, for the 1-bit quantization cannot be assumed to be true. For some systems, and in particular for FMCW radars as will be highlighted in Part II, the coarse acquisition might discard information about the signals of interest that cannot be recovered, regardless from the reconstruction algorithm or the number of measurements used.

Indeed, beside the obvious loss of information regarding the amplitude as $\mathcal{Q}_\epsilon(\boldsymbol{y}) = \mathcal{Q}_\epsilon(c\boldsymbol{y})$ for any $c \in \mathbb{R}_+$, some linear models might suffer from ambiguities where two different vectors are, once measured and quantized, sent to the same 1-bit measurement vectors.

The canonical example of this ambiguity has been introduced by [PV12] for Bernoulli measurements. Considering a measurement matrix $\boldsymbol{\Phi} \in \mathbb{R}^{m\times N}$ made of ± 1, two vectors $\boldsymbol{u} = [1, 0, \dots, 0]^T$ and $\boldsymbol{v} = [1, \alpha, \dots, 0]^T$ give the same 1-bit measurements, with $\alpha < 1$, *i.e.*,

$$\mathcal{Q}_\epsilon(\boldsymbol{\Phi u}) = \mathcal{Q}_\epsilon(\boldsymbol{\Phi v}). \tag{3.6}$$

Indeed, given the distribution of each $\Phi_{ij} \in \pm 1$ with $i \in [m], j \in [N]$, one can directly observe that $\forall i \in m, \operatorname{sign}(\boldsymbol{\Phi u})_i = \operatorname{sign}(\pm 1) = \operatorname{sign}(\pm 1 \pm \alpha) = \operatorname{sign}(\boldsymbol{\Phi v})_i$. The two 1-bit measurements vectors are identical. Chapters 4 and 5 will elaborate more on the existence of such ambiguous scenarios in the radar setting.

One way of solving this issue is to dither the measurements, *i.e.*, adding a variable before the quantization, *i.e.*, for $i \in [m]$,

$$z_i = \mathcal{Q}_\epsilon(y_i + \xi_i).$$

Dithering has been studied in various fields [Wan97]. It is, for example, used in image processing [Buh+98] to perform image compression without the apparition of *colour banding*. This effect is clearly visible in Fig.3.7b where the image in Fig.3.7a is directly quantized to 1-bit. Fig.3.7c uses the same quantization but with an added random dither. Although both images are represented using only binary values, the quantization used in Fig.3.7c is able to visually retain more information about the original image.

In QCS, two modalities exist for the generation of this dither. The dither is either randomly generated [JC17; XJ19] and thus does not depend on the previous measurements, or it is specifically designed at each measurement or estimation [Kam+12; Dir19; Bar+17]. This thesis will focus on the former, as it provides the simplest hardware implementation. Indeed, generating this dither at each measurement requires a communication between the acquisition and the processing of the data, where as a fully random dither does not.

(a) $\boldsymbol{y}$ (b) $\mathcal{Q}_\epsilon(\boldsymbol{y})$ (c) $\mathcal{Q}_\epsilon(\boldsymbol{y}+\boldsymbol{\xi})$

Figure 3.7: 1-bit quantization applied to the Fig.3.1a in grayscale; in (a) the grayscale image; in (b) with the deterministic 1-bit quantization; in (c) with an additive dither added before the quantization).

In [XJ19], the authors proposed and studied uniform reconstruction guarantee when applying a uniform dither whose dynamic matches the resolution of the quantizer. Although their study applies to the more general mid-rise quantizer as introduced in the section above, we adapt and present these results in the context of 1-bit acquisition.

The dithered 1-bit quantizer becomes

$$\mathcal{Q}_\epsilon^+(\boldsymbol{y}) = \mathcal{Q}_\epsilon(\boldsymbol{y}+\boldsymbol{\xi}),$$

with $\boldsymbol{\xi} \in \mathbb{R}^m \sim \mathcal{U}^m(-\frac{\epsilon}{2}, \frac{\epsilon}{2})$ if $\boldsymbol{y}$ is strictly real and $\boldsymbol{\xi} \in \mathbb{C}^m \sim \mathcal{U}^m(-\frac{\epsilon}{2}, \frac{\epsilon}{2}) + \mathrm{i}\mathcal{U}^m(-\frac{\epsilon}{2}, \frac{\epsilon}{2})$ if the measurements to be quantized lie in the complex domain.

Measurements acquired by this process and whose measurement matrix follows the RIP defined in Def. 3.1, can be shown to obey the following theorem in [XJ19]:

Remark 3.3 (Remark from [XJ19] in Sec.7.3.A)**.** If $\boldsymbol{\Phi} \in \mathbb{C}^{m\times N}$ is generated by a RIP matrix distribution and if $\boldsymbol{\xi} \sim \mathcal{U}^m([0,\epsilon])$, then, with high probability and up to some missing log factors, PBP produces for all vectors $\boldsymbol{x} \in \bar{\Sigma}_s^N \in \mathbb{B}^N$ an estimate $\hat{\boldsymbol{x}}$ from $\boldsymbol{y} = \mathcal{Q}_\epsilon^b(\boldsymbol{\Phi}\boldsymbol{x}+\boldsymbol{\xi})$ whose error has the

following decay rate when m increases:

$$\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2 = \mathcal{O}\left(\frac{1+\epsilon}{\sqrt{m}}\right).$$

One can see that now, compared to the ambiguous case presented in (3.6), there is no ambiguous scenario where the reconstruction is bounded. If one wants a better estimate of the sparse vector, it is sufficient to simply increase the number of measurements. Presenting the theory on which this theorem relies is out of the scope of this simple introduction to QCS. However, we highlight here one new property that the quantized measurements $\boldsymbol{z}$ now enjoy thanks to this dither.

Lemma 3.6. *For a vector $\boldsymbol{y} \in \mathbb{C}^m$, if the resolution ϵ is set as $\epsilon \geq 2\|\boldsymbol{y}\|_\infty$, then the quantized and dithered 1-bit measurements follow*

$$\mathbb{E}\{\mathcal{Q}_\epsilon^+(\boldsymbol{y})\} = \boldsymbol{y}.$$

This means that the quantizer $\mathcal{Q}_\epsilon^+(\cdot)$ can potentially recover all of the information about $\boldsymbol{y}$. Fig. 3.8 gives some insights on the underlying effect of the added dither. It is interesting to note that this effect heavily relies

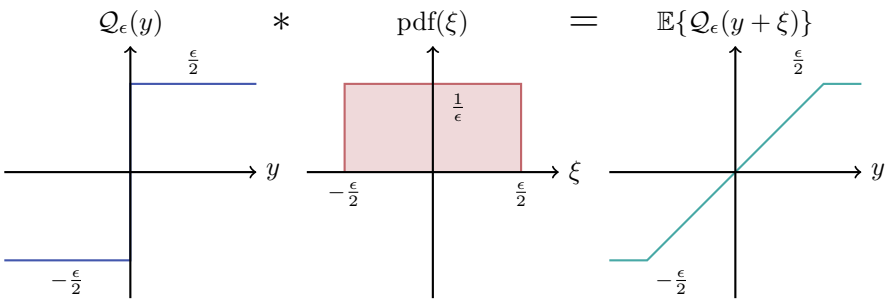


Figure 3.8: Graphical representation of the effect of the dither on the 1-bit quantization.

on the dynamic of the dither ϵ with respect to the dynamic of the considered measurements. The next chapter will study in depth the effect and requirements that this additive dithering has on radar signal processing.

Part II

Additive Dithering

Chapter 4

Range estimation using an FMCW radar

IN this chapter, we investigate a trade-off between the number of radar observations (or measurements) and their resolution in the context of radar range estimation. To this end, we introduce a novel estimation scheme that can deal with strongly quantized received signals, going as low as 1-bit per signal sample. We leverage for this a dithered quantized compressive sensing framework that can be applied to classic radar processing and hardware. This allows us to remove ambiguous scenarios prohibiting correct range estimation from (undithered) quantized base-band radar signal. Two range estimation algorithms are studied: Projected Back Projection (PBP) and Quantized Iterative Hard Thresholding (QIHT). The effectiveness of the reconstruction methods combined with the dithering strategy is shown through Monte Carlo simulations. Furthermore we show that: *(i)*, in dithered quantization, the accuracy of target range estimation improves when the bit-rate (*i.e.*, the total number of measured bits) increases, whereas the accuracy of other undithered schemes saturate in this case; and *(ii)*, for fixed, low bit-rate scenarios, severely quantized dithered schemes exhibit better performances than their full resolution counterparts. These observations are confirmed using real measurements obtained in a controlled environment, demonstrating the feasibility of the method in real ranging applications.

4.1 Introduction

Civilian radar applications such as automotive radar design or the growing field of smart cities are more and more in need of small form factor and affordable radars [HS09; End13; CE18]. As these complex applications often require the deployment of many radar sensors working in a collaborative mode, the increasing amount of data recorded by these systems challenges both data transmission and processing techniques.

In this chapter we investigate a trade-off between the number of radar observations (or measurements) and their resolutions in the case of an FMCW radar with one transmitting and one receiving antenna. While this setting might seem restrictive as it only considers target range recovery, its set-up allows us to perform thorough tests using both simulations and actual radar measurements. To this end, two range estimation algorithms, adapted to quantized radar signal, are used: Projected Back Projection (PBP) [XJ19] and Quantized Iterative Hard Thresholding (QIHT) [JDD13]. Compared to [Feu+18a], this work deeply investigates the comparisons between severely quantized and high-resolution measurements constrained to the same bit-rate, *i.e.*, between quantity and quality.

One-bit quantized compressive radar schemes have been studied in, *e.g.*, [DZ15; Li+16; Wan+19]. One limiting effect is that, as the digitization becomes coarser, ambiguities might appear between different unquantized signals — and thus different target configurations — that are digitized to the same bits, rendering the estimation ambiguous. These works, however, failed to address these ambiguities. Our previous work in 1-bit quantization applied to Frequency-Modulated Continuous-Wave radar (FMCW) [Feu+18a] showed that these ambiguities do happen in realistic settings and measurements and can be counteracted using a pre-quantization *dither*. Dithering amounts to adding a designed noise on the signal, before quantizer's action, with the goal of attenuating quantization distortions [GS93; XJ19]. This procedure is also used in, *e.g.*, LIDAR imaging [RDG18] where dithering is implemented in a real set-up by physically varying a time-delay before the acquisition, and was studied for high sampling rate ADCs [Bra96].

Let us summarize the main contributions of this work: *(i)* we show that ambiguities due to the combination of the intrinsic radar Fourier domain

with harsh quantization exist and are removed using dithered quantization; *(ii)* we observe that as the number of measurements m grows, non-dithered quantization yields range estimation error (using either PBP or QIHT) that saturates whereas the dithered schemes reach a decaying error when m increases; *(iii)* we show that QIHT provides the best performances at low resolution for harsh bit-rate condition; and *(iv)* we confirm all the above observations on a controlled laboratory set-up using an FMCW radar.

The rest of the chapter is structured as follows. In Sec. 4.2, the complete FMCW radar model (*i.e.*, its transmission and reception principles) is introduced, as well as a linear inverse problem formulation focused on a Fourier sensing model of the range profile. Sec. 4.3 defines the quantization procedure applied on the received radar signal. We then prove that unavoidable ambiguities are induced by this scheme, *i.e.*, the existence of distinct received signals (and thus distinct range profiles) whose quantized measurements are identical. A dithered quantizer is then proposed to cancel out these ambiguous situations. Sec. 4.4 describes two algorithms capable to estimate sparse range profiles from quantized observations, namely PBP and QIHT. Finally, we demonstrate the efficiency of our approach through intensive Monte Carlo simulations in Sec. 4.5, and via real radar measurements in Sec. 4.6, before concluding in Sec. 4.7.

4.2 Radar System Model

This work tackles the issue of quantizing FMCW radar signals and studies its effect on the localization performance. The time domain expression of the received signal coming from a target located at a range R has been introduced in (2.9). For this signal to be processed it first needs to be sampled. Sampling $r(t)$ at the receiver at a rate T/N for some integer N, *i.e.*, at time samples $t_i := i(T/N), i \in \mathbb{Z}$, gives

$$r[i] = A \exp\big(- \mathsf{i}\, 2\pi f_i \tfrac{2R}{\mathsf{c}}\big), \tag{4.1}$$

with $f_i := f_c(t_i) = f_0 + B\,(\frac{i}{N} \bmod 1)$. A single ramp can thus be sampled over at most N time samples, which implicitly determines both the resolution $\mathsf{c}/(2B)$ and the maximum range $R_{\max} := \mathsf{c}N/(2B)$ at which R can be estimated.

Let us now turn to a multi-target scenario restricted to a purely additive model; all the targets are in a direct line of sight from the radar, without any possible multipath propagation. Taking into account the radar range resolution ($\mathsf{c}/2B$) and $R_{\max}$, we discretize the range domain $(0, R_{\max}]$ with N ranges $\mathcal{R} := \{R_n := n(\mathsf{c}/2B), 1 \leq n \leq N\}$. A range profile resulting from s targets with ranges in $\mathcal{R}$ is expressed as a s-sparse vector $\boldsymbol{a} = (a_1, \cdots, a_N)^\top$, *i.e.*, the amplitude $a_n \neq 0$ if there is a target at the n^{th} range bin R_n, and $\|\boldsymbol{a}\|_0 := |\operatorname{supp}(\boldsymbol{a})| \leq s$. Then, the single target case (4.1) generalizes to the multi-target sensing model

$$r[i] = \sum_{n=1}^{N} a_n e^{-\mathrm{i}\, 2\pi f_i \frac{2R_n}{\mathsf{c}}} = \sum_{n=1}^{N} a'_n e^{-\mathrm{i}\, 2\pi \frac{in}{N}}, \tag{4.2}$$

where $a'_n = a_n e^{-\mathrm{i}\, 2\pi n f_0/B}$. In words, each observation $r[i]$ at time t_i amounts to probing the i^{th} frequency of the discrete Fourier transform the range profile $\boldsymbol{a}' = (a'_1, \cdots, a'_N)^\top$. Hereafter, since $\boldsymbol{a}'$ encodes the same range profile as $\boldsymbol{a}$ (up to a modulation), we drop the prime symbol for the sake of simplicity. Classically, in a Nyquist sensing scenario, if we collect N samples $\boldsymbol{r} = (r[1], \cdots, r[N])^\top$, (4.2) is equivalent to $\boldsymbol{r} = \boldsymbol{F}^* \boldsymbol{a}$, with $\boldsymbol{F}$ the Fourier matrix (*i.e.*, $F_{in} := \exp(\mathrm{i}\, 2\pi \frac{in}{N})$), and an inverse Fourier transform recovers $\boldsymbol{a}$. For noisy observations, a sampling over multiple ramps — hence reaching an oversampled sensing model — yields a robust estimate of $\boldsymbol{a}$.

In this work, we leverage the sparsity assumption made on $\boldsymbol{a}$ to allow this estimation through severely quantized, possibly oversampled, received signal samples. Without quantization, Compressive Sensing (CS) theory from partial random Fourier sensing matrices shows that, with high probability, we can recover any s-sparse vector $\boldsymbol{a}$ from only $m = O(s \log^4(N))$ random samples of $\boldsymbol{r}$ [FR13]. However, as made clear in Sec. 4.3, QCS aims to reduce the impact of signal measurement quantization in signal estimation by possibly increasing the number of measurements beyond N; what truly matters in QCS is indeed the total bit-rate $\mathcal{B}$ (*i.e.*, $m\times$ the bit depth b) used to encode the observations [Bou+15; XJ19].

Consequently, our sensing scheme is determined by sampling the received signal $r(t)$ over a set of m (discrete) time samples $\mathcal{T} = \{t'_i : 1 \leq i \leq m\}$ determined as follows. If $m < N$, then $(t'_1, \cdots, t'_m)$ is picked uniformly at random among all possible subsets of m time samples of $\{t_i : 1 \leq i \leq N\}$. If $m > N$, in an effort to obtain an acquisition time as short as possible,

we take then $t'_i = t_i$ for $1 \leq i \leq N\lfloor m/N \rfloor$, *i.e.*, the first $\lfloor m/N \rfloor$ ramps are fully sampled, and the set of $m' = m - N\lfloor m/N \rfloor$ remaining samples is picked uniformly at random among all possible subset of m' time samples of $\{t_i : N\lfloor m/N \rfloor + 1 \leq i \leq N(\lfloor m/N \rfloor + 1)\}$, *i.e.*, the last ramp is randomly sub-sampled.

Correspondingly, these time samples are associated with m, possibly non-distinct, frequencies $\{f'_i = f_c(t'_i) : 1 \leq i \leq m\}$. Finally, if Ω is a multiset (*i.e.*, a set with repeated elements) representing the indices of these frequencies in $[N]$, the final CS model, before quantization, reads

$$\boldsymbol{r} = \boldsymbol{\Phi}\boldsymbol{a} = \boldsymbol{F}^*_\Omega \boldsymbol{a}, \tag{4.3}$$

where $\boldsymbol{\Phi} := \boldsymbol{F}^*_\Omega$, $\boldsymbol{F}_\Omega$ gathers the (possibly repeated) columns of $\boldsymbol{F}$ indexed in Ω, and $\boldsymbol{r}$ follows the sampling of $r(t)$ over $\mathcal{T}$. Note that for $m > N$, the addition of a dither ensures that the observations of r over repeated frequencies carry additional information (see Sec. 4.3).

4.3 Quantization: Model & Ambiguity

We select in this work on a uniform b-bit scalar quantizer applied componentwise onto complex vectors as defined in 3.3, separately on the real and the imaginary domains, *i.e.*,

$$\mathcal{Q}^b_\epsilon(\boldsymbol{r}) = \mathcal{Q}^b_\epsilon(\Re(\boldsymbol{r})) + \mathsf{i}\, \mathcal{Q}^b_\epsilon(\Im(\boldsymbol{r})), \tag{4.4}$$

where b is the number of bits per vector component (*i.e.*, the I and Q channels), or *bit depth*. This quantization takes place on the received baseband signal $\boldsymbol{r}$ using ADCs with a resolution of b bits. In (4.4), $\mathcal{Q}^b_\epsilon(\cdot)$ is the standard mid-riser quantizer of quantization step size $\epsilon > 0$ [GS93; XJ19]

$$\mathcal{Q}^b_\epsilon(\lambda) := \epsilon \lfloor \frac{\lambda}{\epsilon} \rfloor + \frac{\epsilon}{2}, \quad \forall \lambda \in \mathbb{R}.$$

The step size is set to $\epsilon = \alpha_b \Delta$, where Δ is the dynamic range of the ADC, *i.e.*, its voltage range $[-\Delta, \Delta]$, and $\alpha_b = 2^{1-b}$ ensures that the bit-depth of each sample is b. For example, for $b = 1$, the ADC is then a simple voltage comparator over its domain, *i.e.*, $2\mathcal{Q}^1_\Delta(\cdot)/\Delta \equiv \mathrm{sign}(\cdot)$. This

definition assumes that the quantizer is adjusted to the variations of $\boldsymbol{r}$, *i.e.*, we must have $\Delta \geq \|\boldsymbol{r}\|_\infty$, with Δ as small as possible to minimize the quantization distortion which scales like $O(\epsilon)$. Note that one can also decide to set $\Delta \geq |r[i]|$ only for a significant fraction of indices i, *e.g.*, if $\|\boldsymbol{r}\|_\infty$ is not bounded or if we allow for some saturation. Hereafter, we just assume that Δ is given.

Similarly to Section 3.3.2, let us stress an important limitation of a too direct quantization of the radar sensing model (4.3): the existence of distinct vectors whose quantized Fourier observations are sent to the same quantized vector, rendering the estimation process ambiguous. This bears similarities with known ambiguities in 1-bit CS with binary matrices [PV14] and for QCS for multiple antennas and a single target [Feu+18a] (see Chapter 5). We show here that the same effect exists for multiple targets and one receiving antenna.

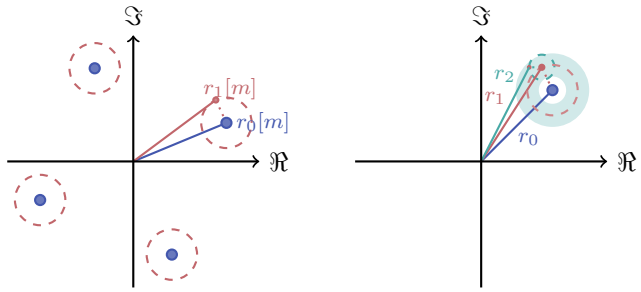
This ambiguity is explained by the following construction. Given two distinct $n_0, n_1 \in [N]$, we build $\boldsymbol{a}_0 = \boldsymbol{b}_{n_0} e^{-\mathrm{i}\,\psi_{n_0}}$ and $\boldsymbol{a}_1 = \boldsymbol{a}_0 + \gamma \boldsymbol{b}_{n_1} e^{-\mathrm{i}\,\psi_{n_1}}$, with ψ_{n_0} and ψ_{n_1} two arbitrary phases in $[-\pi, \pi)$, $0 < \gamma < 1$, and $\boldsymbol{b}_i \in \{0,1\}^N$ the (canonical) vector whose components are all 0 but the i^{th} ($i \in [N]$). The signal $\boldsymbol{a}_0$ can be seen as one unit-amplitude target at location R_{n_0}, while $\boldsymbol{a}_1$ contains an additional target at R_{n_1} with amplitude γ. According to the CS model (4.3), the acquired received signals are $\boldsymbol{r}_0 = \boldsymbol{\Phi}\boldsymbol{a}_0$ and $\boldsymbol{r}_1 = \boldsymbol{\Phi}\boldsymbol{a}_1$, with

$$\begin{aligned} r_0[i] &= e^{-\mathrm{i}\,\psi_{n_0}} e^{-\mathrm{i}\,2\pi \frac{in_0}{N}}, \\ r_1[i] &= e^{-\mathrm{i}\,\psi_{n_0}} e^{-\mathrm{i}\,2\pi \frac{in_0}{N}} + \gamma e^{-\mathrm{i}\,\psi_{n_1}} e^{-\mathrm{i}\,2\pi \frac{in_1}{N}}, \end{aligned} \tag{4.5}$$

Interestingly, there exist parameter values where the quantizer (4.4) sends the two signals to the same quantized vector, *i.e.*, for which the ambiguity condition (AC) holds:

$$\mathcal{Q}_\epsilon^b(\boldsymbol{r}_0) = \mathcal{Q}_\epsilon^b(\boldsymbol{r}_1). \tag{AC}$$

Consequently, in these cases, while the ℓ_2-distance $\|\boldsymbol{a}_1 - \boldsymbol{a}_0\| = \gamma$ is nonzero, recovering both $\boldsymbol{a}_1$ and $\boldsymbol{a}_0$ from their identical quantized observations is impossible. Let us study when (AC) occurs for 1-bit quantization ($b = 1$), *i.e.*, $\mathcal{Q}_1^{\mathbb{C}}(\cdot) \propto \operatorname{sign}(\Re(\cdot)) + \mathrm{i}\operatorname{sign}(\Im(\cdot))$. In this case, (AC) involves that $r_0[i]$ and $r_1[i]$ are always in the same quadrant of the complex plane $\mathbb{C}$ for all i. Since from (4.5) $r_1[i]$ lies on a circle of center $r_0[i]$ and radius γ in $\mathbb{C}$,



(a) (b)

Figure 4.1: test (a) Graphical representation of r_0 and r_1 and the domain on which $\boldsymbol{r}_1$ lies. (b) Extension to 3 targets, where the domain to consider for the verification of (AC) is enlarged.

regardless of the values of ψ_{n_1} or R_{n_1} (see Fig. 4.1a), (AC) holds if

$$\min_i \min(|\Re(r_0[i])|, |\Im(r_0[i])|) \ > \ \gamma. \tag{4.6}$$

As $r_0[i] = e^{-\mathrm{i}(\psi_0 + 2\pi \frac{n_0 i}{N})}$, (4.6) shows a clear dependency between the parameters ψ_0, N, m, and n_0 for two quantized vectors to be indistinguishable. For instance, if $n_0 = N/4$, then we just need $\gamma < \min(|\sin\psi_0|, |\cos\psi_0|)$ for (AC) to hold for any values of ψ_1 and n_1 (see Fig. 4.1a). Similar examples can be constructed for other values of n_0, as well as with multiple targets, with then more restriction on the amplitudes of the additional targets, as suggested in Fig. 4.1b. For $b > 1$, there also exist vectors satisfying (AC), but their ℓ_2-distance must decay if b increases since $\mathcal{Q}_b$ splits $\mathbb{C}$ into square cells of size $2^{1-b}\Delta$. Therefore, if an algorithm wrongly estimates $\boldsymbol{r}_1$ with the value of $\boldsymbol{r}_0$, its error decays as 2^{-b} if b increases, but this error is not ensured to decay if m increases.

In this work, we stress that the previous ambiguities can be removed by deliberately introducing randomness in the quantization, *i.e.*, by inserting a random dither in the quantizer input. Consequently, one can design algorithms whose estimation error of range profile decay as m increases. While dithered quantization is a well-known strategy to improve signal estimation techniques (see, *e.g.*, [GS93; Llo52; RDG18]), its use in quantized compressive sensing is recent and we follow here the approach of [XJ19].

Given a range profile $\boldsymbol{a} \in \mathbb{C}^N$, our dithered QCS sensing model is thus defined by

$$\boldsymbol{z} = \mathcal{Q}_\epsilon^b(\boldsymbol{\Phi a} + \boldsymbol{\xi}), \tag{4.7}$$

where $\boldsymbol{\xi} \in \mathbb{C}^m$ is a complex dither defined as $\xi_i = \xi_i^{\Re} + \mathsf{i}\,\xi_i^{\Im}$, with $\xi_i^{\Re}, \xi_i^{\Im} \sim \mathcal{U}(-\frac{\epsilon}{2}, \frac{\epsilon}{2})$. This dither induces more diversity in the quantized measurements, especially for $m > N$. Moreover similarly to Lemma 3.6, $\mathbb{E}_\xi\{\mathcal{Q}_\epsilon^b(\boldsymbol{\Phi a} + \boldsymbol{\xi})\} = \boldsymbol{\Phi a}$, *i.e.*, the dither cancels out the quantization error in expectation, or, equivalently, if m is large [XJ19]. Note that this also changes the dynamic range of the signal before quantization, *i.e.*, we must adapt the range $\Delta \geq \|\boldsymbol{r}\|_\infty + \frac{\epsilon}{2}$.

4.4 Reconstruction Algorithm

To reconstruct the range profile $\boldsymbol{a}$ from the quantized measurements $\boldsymbol{y}$, two algorithms are studied. The first is Projected Back Projection (PBP) that was introduced in Chapter 3 in Def. 3.3 and is defined as follow:

$$\hat{\boldsymbol{a}} = \tfrac{1}{m}\mathsf{H}_s\big(\boldsymbol{\Phi}^H \boldsymbol{z}\big),$$

where s is the range profile sparsity, assumed known a priori, H_s is the hard-thresholding operator setting all the components of its vector input to zero but those with the s largest amplitudes.

The advantages of PBP are threefold. First, its complexity is $O(N \log N)$ since $\boldsymbol{\Phi}^H$ only requires the computation of an inverse FFT applied on a zero-padding[1] of $\boldsymbol{z}$ from Ω to $[N]$ (or $[\rho N]$ for $\rho = O(1)$ ramps) and H_s involves a vector component ordering of $O(N \log N)$ computations. Second, as a function of $\boldsymbol{z}$, PBP does not explicitly invoke the dither $\boldsymbol{\xi}$; its implementation only requires the knowledge of $\boldsymbol{\Phi}$, *i.e.*, of Ω. Finally, in the context of dithered QCS, PBP enjoys a reconstruction error that decays when m increases for all sensing matrices $\boldsymbol{\Phi}$ respecting with high probability the restricted isometry property (RIP) [XJ19], such as for the random partial Fourier matrix in (4.3). Remark 3.3 from [XJ19] shows that, for a sparse

[1] In this sense, PBP is similar to a *Maximum Likelihood Estimator*.

range profile $\boldsymbol{a}$, the reconstruction guarantees is

$$\|\boldsymbol{a} - \hat{\boldsymbol{a}}\| = \mathcal{O}(m^{-\frac{1}{2}}). \tag{4.8}$$

In other words, compared to the undithered context, no counterexamples exist that would make this error stagnate when m is increased.

Note that (4.8) is a root-mean-square error bound for the estimation of $\boldsymbol{a}$. In this work, our interest is, however, to characterize the range recovery of several targets, *i.e.*, the support of $\boldsymbol{a}$. Interestingly, since $\|\boldsymbol{a}-\hat{\boldsymbol{a}}\|_\infty \leq \|\boldsymbol{a}-\hat{\boldsymbol{a}}\|$ one can show the following.

Lemma 4.1 (Support recovery)**.** *If $\boldsymbol{a}$ is s-sparse, with s given, and if we know that $\min\{|a_i| : i \in \operatorname{supp}\boldsymbol{a}\} > \eta$ for some $\eta > 0$, then, one can expect that*

$$m \geq C/\eta^2 \quad \Rightarrow \quad \operatorname{supp}(\hat{\boldsymbol{a}}) = \operatorname{supp}(\boldsymbol{a}), \tag{4.9}$$

for some $C > 0$.

Proof. Indeed, support recovery is ensured if $|\hat{a}_i| > |\hat{a}_j|$ for all $i \in \operatorname{supp}\boldsymbol{a}$ and all $j \in [N] \setminus \{\operatorname{supp}\boldsymbol{a}\}$, which is achieved if $|a_i| - |\hat{a}_i - a_i| > |\hat{a}_j - a_j|$. This holds if $|a_i| > \eta > 2\|\boldsymbol{a} - \hat{\boldsymbol{a}}\|_\infty = O(m^{-1/2})$, or if $m \geq C/\eta^2$. □

While requiring a single iteration, PBP does not ensure that its estimate $\hat{\boldsymbol{a}}$ is *consistent* with $\boldsymbol{z}$, *i.e.*, $\mathcal{Q}_\epsilon^b(\boldsymbol{\Phi}\hat{\boldsymbol{a}} + \boldsymbol{\xi}) \neq \boldsymbol{z} = \mathcal{Q}_\epsilon^b(\boldsymbol{\Phi}\boldsymbol{a} + \boldsymbol{\xi})$; the quantized sensing model is thus not fully exploited while estimating $\boldsymbol{a}$ from $\boldsymbol{z}$. To solve this situation, [JDD13] has proposed the Quantized Iterative Hard Thresholding (QIHT) algorithm, *i.e.*, a variant of the Iterative Hard Thresholding (IHT) [BD09] (see Def. 3.4) and of the Binary IHT [Jac+13], iteratively enforcing both consistency and sparsity of a signal estimate. QIHT is defined by

$$\hat{\boldsymbol{a}}^{j+1} = \mathsf{H}_s\big[\hat{\boldsymbol{a}}^j + \tfrac{\mu}{M}\boldsymbol{\Phi}^H\big(\boldsymbol{z} - \mathcal{Q}_\epsilon^b(\boldsymbol{\Phi}\hat{\boldsymbol{a}}^j + \boldsymbol{\xi})\big)\big], \tag{4.10}$$

where j is the iteration index, μ is a step size parameter, and $\hat{\boldsymbol{a}}^0$ is the PBP estimate. Compared to PBP, this algorithm is not ensured to converge. However, numerically, QIHT often provides a sparse and consistent estimate. If this happens at the J^{th} iteration, *i.e.*, $\boldsymbol{z} - \mathcal{Q}_\epsilon^b(\boldsymbol{\Phi}\hat{\boldsymbol{a}}^J + \boldsymbol{\xi}) = \boldsymbol{0}$,

and if $\boldsymbol{\Phi}$ is a random Gaussian matrix, the QIHT estimate $\hat{\boldsymbol{a}} = \hat{\boldsymbol{a}}^J$ reaches an error $\|\boldsymbol{a} - \hat{\boldsymbol{a}}\| = O(1/m)$ [Jac16; Fri+20]. Consequently, we decide to also investigate the efficiency of QIHT for the radar sensing model (4.7).

While QIHT has more to offer in terms of reconstruction by enforcing the consistency, one must also note that knowing the dither at the reconstruction, as imposed by the computation of $\mathcal{Q}_\epsilon^b(\boldsymbol{\Phi}\hat{\boldsymbol{a}}^J + \boldsymbol{\xi})$ in (4.10), will impact the physical implementation of the system. Indeed PBP could use analogical random noise source such as a noise diode [Bra96], whereas QIHT would require a more advanced implementation. A more in-depth discussion of this different requirements is provided in Chapter 7.

4.5 Numerical Results

We here challenge the possibility of recovering sparse range profiles from quantized radar observations, *i.e.*, from measurements associated with the dithered QCS model (4.7). To this aim, we present the result of extensive Monte Carlo (MC) simulations for various parameters of our setup: we have set the sparsity level s — the number of targets — in $[2, 10]$, a total bit-rate $\mathcal{B} = bm$ in $[2^3, 2^{13}]$ with measurement number m in $[2^3, 2^{13}]$ and a bit depth $b \in [1, 32]$, $N = 256$. Concerning QIHT, we have set $\mu = 1$ and a total number of iterations between 20 and $100s$, with an early stop if, either, the *consistency* $m^{-1} \sum_k (\mathcal{Q}_\epsilon^b(\boldsymbol{\Phi}\hat{\boldsymbol{a}}^j + \boldsymbol{\xi})_k = z_k)$ between $\hat{\boldsymbol{a}}^j$ and $\boldsymbol{a}$ exceeds 95%, or if the consistency decreases from the previous iterations. Note that unquantized observations are actually associated with 32-bits floating point variables.

For any fixed values of these parameters, 2000 trials of the MC simulations were considered by randomly drawing both the sparsity range profile $\boldsymbol{a}$, the radar sensing matrix $\boldsymbol{\Phi}$, and the dither $\boldsymbol{\xi}$. The resulting full resolution signals $\boldsymbol{\Phi a}$ were then dithered and quantized from (4.7) before estimation of $\boldsymbol{a}$ from PBP or QIHT.

More precisely, each s-sparse vector $\boldsymbol{a}$ was randomly built by picking its support uniformly at random among the $\binom{N}{s}$ possible supports, and by independently drawing its s non-zero components as the random variable $C \exp(\mathrm{i}\,\psi)$, with $C \sim \mathcal{U}([0, 1])$ and $\psi \sim \mathcal{U}([0, 2\pi))$, before the normalization $\boldsymbol{a} \leftarrow \boldsymbol{a}/\|\boldsymbol{a}\|_\infty$. Following Sec. 4.2, the random sensing matrix $\boldsymbol{\Phi} = \boldsymbol{F}_\Omega^*$ was

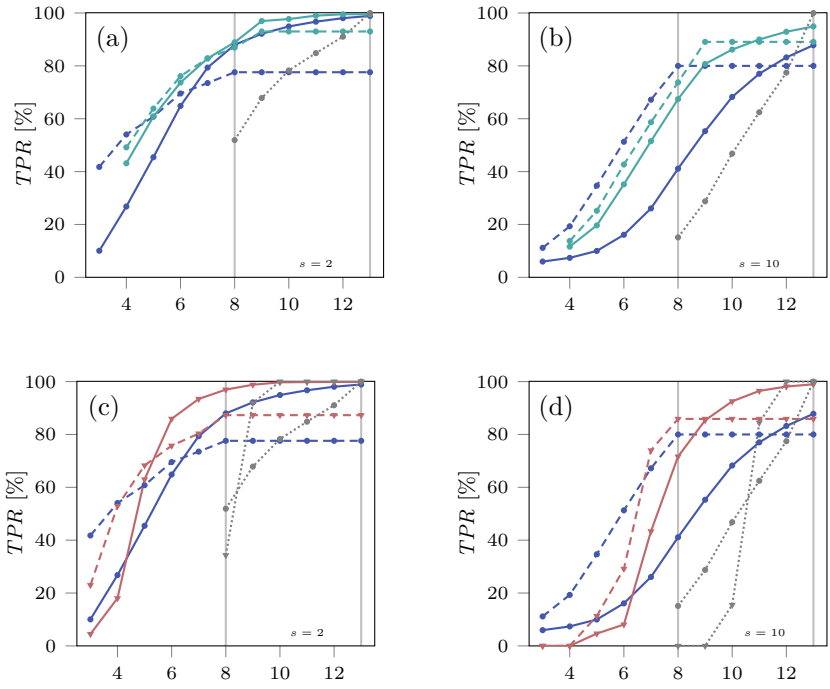


Figure 4.2: [best viewed in color] (a) and (b): TPR vs $\log_2 \mathcal{B}$ for PBP; (c) and (d): Comparison between PBP (disks) and QIHT (triangles) in function of $\log_2 \mathcal{B}$. In all figures, solid, dashed and dotted curves stand for dithered, undithered and unquantized schemes, respectively. The first (second) **gray** vertical line represents a bit-rate of 2^8 (2^{13}) bits corresponding to $m = 256$ ($m = 8192$) for 1-bit and $m = 16$ ($m = 256$) for no quantization. In (a) and (b), the resolution is represented by colors, **blue** for 1-bit, **green** for 2-bits and **gray** in absence of quantization. In (c) and (d) **blue** stands for 1-bit PBP, **red** for 1-bit QIHT and **gray** for no quantization. Figures (a,c) and (b,d) are for $s = 2$ and $s = 10$, respectively.

generated according to a random draw of $\mathcal{T}$ (inducing a random multiset Ω). The complex dither $\boldsymbol{\xi}$ was generated as a complex random uniform vector adjusted to $\mathcal{Q}$ and $\epsilon = \alpha_b \Delta$ (see Sec. 4.3).

We assessed the efficiency of the range profile estimation by measuring the accuracy of the support recovery. In particular, we computed the True Positive Rate, *i.e.*, TPR = TP/s, where the number of True Positives TP is the number of estimated targets that were actually part of the true range profile, *i.e.*, TP := $|\operatorname{supp}(\hat{\boldsymbol{a}}) \cap \operatorname{supp}(\boldsymbol{a})|$. We solely focus on the TPR as

opposed to also including the False Positive Rate as the number of targets (*i.e.*, s) is known a priori.

As a first evaluation of the potential of dithered quantization, we have focused on the performances of PBP; hence establishing a reference level for further experiments. Fig. 4.2a shows that for $\mathcal{B} = bm \in [2^8, ..., 2^{13}]$, low bit-depth strategies (*e.g.*, $b \in 1, 2$) outperform the TPR of high-resolution quantizers (with $m \leq N$ at $b = 2^5$ in this bit-rate range). Moreover, in Fig. 4.2a as well as in Fig. 4.2b, we clearly observe a TPR saturation for undithered schemes from $m = \mathcal{B}/b \geq N$, *i.e.*, for $\mathcal{B} \geq 2^8$ and $\mathcal{B} \geq 2^9$ at one and two-bit quantization, respectively; deterministic quantization does not provide more information from repeated quantized measurements in our synthetic examples. The TPR performances of the dithered schemes, however, continue to scale as $\mathcal{B}$ increases, as hinted by (4.8) and (4.9).

The simplicity of the PBP algorithm, *i.e.*, the absence of an explicit usage of the dither and the non-consistency of the produced estimate (see Sec. 4.4), limits its ability to distinguish targets with weak amplitudes before the quantization level. Therefore, as observed by comparing the TPR of Fig. 4.2a ($s = 2$) and Fig. 4.2b ($s = 10$) for low resolutions dithered schemes, the PBP performances are rather poor for larger values of s at identical values of b and $m = \mathcal{B}/b$. We have thus compared the performances of QIHT — which targets consistency and explicitly uses the dither — and PBP in Fig. 4.2c and Fig. 4.2d in the context of 1-bit quantization, as well as in absence of quantization. In this last case, QIHT and PBP reduces to the IHT and Thresholding algorithms [BD09; FR13], respectively, and IHT also outperforms the Thresholding algorithm by fully exploiting the RIP of $\mathbf{\Phi}$ [FR13]. In these two figures, the TPR of QIHT clearly exceeds the one of PBP in every quantization and bit-rate scenarios. Furthermore, for large values of s, the drop in performances in Fig. 4.2d between the non-dithered and dithered schemes for the 1-bit PBP is reduced for 1-bit QIHT. In Fig. 4.2d, the dithered 1-bit QIHT is markedly better than any other methods for $\mathcal{B} = 2^9$ bits and above, reducing the bit-rate by as much as 93.75% compared to the classic high resolution Nyquist sampling scheme. This bit-rate corresponds to $m \geq 2N = 512$ for 1-bit and $m \geq N/2^4 = 16$ in absence of quantization; at harsh bit-rates quantity outweighs quality.

Finally, we study in Fig. 4.3 the TPR of PBP and QIHT vs s for a

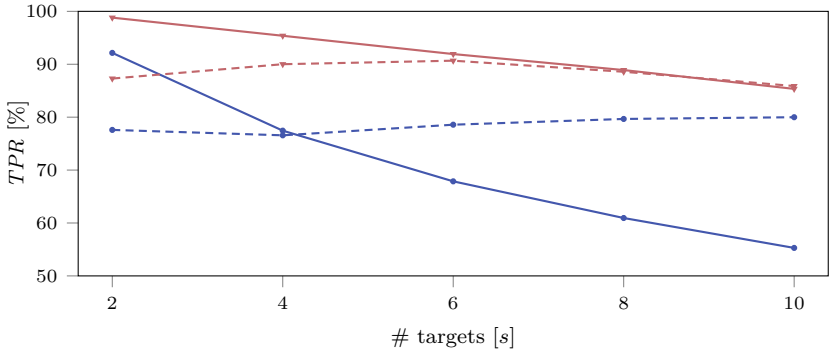


Figure 4.3: [best viewed in color] *TPR* vs number of targets for 1-bit PBP and 1-bit QIHT with $\mathcal{B} = 2^9$ bits, PBP is represented by disks and QIHT by triangles, **blue** stands for 1-bit PBP, **red** for 1-bit QIHT, the solid lines are with additive dithering, the dashed are without dithering.

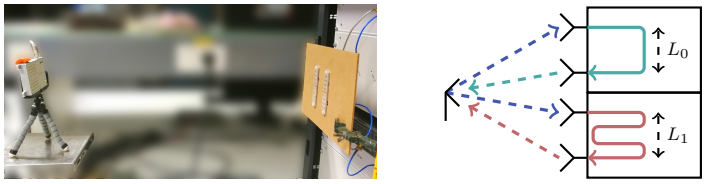
fixed bit-rate $\mathcal{B} = 2^9$. Here also, the gain offered by the explicit knowledge of the dither in QIHT is quite obvious. For low s, both of the dithered schemes have better TPR than their non dithered counterparts. However, as s increases above 4, the performances of the 1-bit dithered PBP plummets quickly below its non dithered version. On the other hand, QIHT with dithering always outperforms its performances with non dithered quantization. We thus conclude that, provided a uniform dithering can be implemented efficiently and later reproduced in QIHT, dithered quantization has always a positive impact on the range estimation.

4.6 Measurements in Laboratory

Sec. 4.5 has focused on the study of range estimation performances from noiseless and synthetic simulations, under a perfect linear sensing model (before quantization) where an idealized radar interacts with point-like targets. We thus present here different tests of resilience of both our model and algorithms by confronting them with real data acquired in a controlled laboratory setting.

The radar used for this experiment is the *KMD2* radar [RFB], *i.e.*, an FMCW radar with one transmitting antenna and three receiving antennas. The radar lies in the K-band and its bandwidth can be extended up to

770 MHz. The AMG43-007 [AMG] is a target simulator distributed by AMG-microwave which is able to simulate a target with varying velocity, range, and power. In the context of this work, two simulators are used with the velocity set to zero and with a power changing according to a logarithmic uniform distribution. This setup allows one to simulate target ranges up to 64 m by 1 m step. We thus set the bandwidth of the FMCW radar to 150 MHz to match this spatial resolution.



(a) (b)

Figure 4.4: (a) Experimental setup: radar in front of the simulator. (b) Block representation of the 2 targets simulator by *AMG*.

The radar is placed in front of these two simulators and emits the frequency pattern (2.3), the signal received by the two simulators is then delayed and attenuated according to user-defined parameters and then re-emitted towards the radar (see Fig. 4.4a and Fig. 4.4b). This process allows the simulation of a specific support while adding the concrete effect associated with the radar that are not taken into account in the developed model. These effects range from the inherent noise in RF and electronics hardware, IQ imbalance, non linearities in the coherent demodulation and all other non idealities related to radar applications (*e.g.*, the non-idealities listed at the end of Chapter 2). This experimental setup has thus the ability to combine the rigor and completeness of Monte Carlo simulations with the possibility to program and repeat specific scenarios (*i.e.*, specific $\boldsymbol{a}$), and to test them against a real acquisition system.

We recorded 196 runs with different sparse range profiles using the same parametrization as in Fig. 4.2c. We observed that the SNR of the configuration in Fig. 4.4a is sufficiently high to neglect the impact of the noise on the quantization. Note that this effect was briefly addressed in [Feu+18a] by experimentally adjusting the dither to the noise amplitude, and its thorough theoretical study is ongoing.

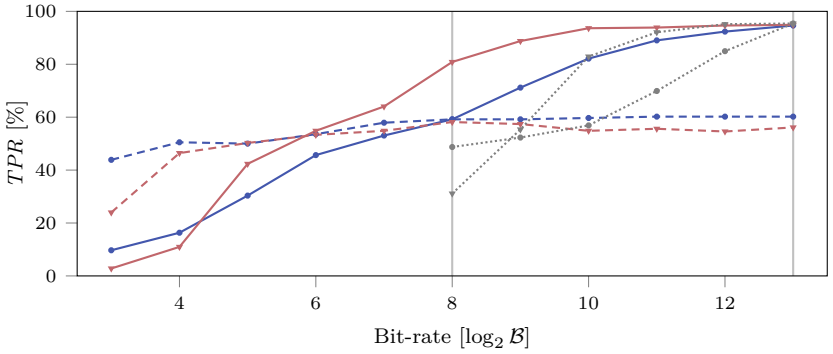


Figure 4.5: [best viewed in color] TPR vs bit-rate using real FMCW radar measurements for $s = 2$. In all figures, PBP is represented by disks and QIHT by triangles, **blue** stands for 1-bit PBP, **red** for 1-bit QIHT, and **gray** for no quantization, the solid lines are with additive dithering, the dashed are without dithering.

The curves in Fig. 4.5 exhibit the same tendencies as in Fig. 4.2c. The only difference is the TPR at which the non-dithered schemes saturate; an effect most probably due to some discrepancies in the range profile amplitudes between this setup and the previous simulations. Once again, 1-bit dithered QIHT is the algorithm with the highest TPR from $\mathcal{B} = 2^6$ bits to 2^{13}, *i.e.*, the bit-rate of a full resolution Nyquist sensing. These results from real measurements are fully consistent with the previously developed theory and simulations; this paves the way to more complete and practical realizations of the proposed quantized architecture.

4.7 Discussion

In this chapter, we demonstrated that a pre-quantization dither removes unavoidable range estimation ambiguities when one quantizes the received radar signal. Moreover, in this dithered scheme, we proved that severe quantization, as low as 1-bit per received signal sample, still allows for an accurate range profile estimation as soon as the total bit-rate is large enough; a tradeoff between the number of radar observations (or measurements) and their resolution (or bit-depth) must be respected. Moreover, we showed that for low bit-rate scenarios, low bit-depth exhibits better performances than an unquantized scheme. These results are achieved thanks to two QCS re-

construction algorithms, PBP and QIHT, that leverage the sparsity of the range profile. Moreover, when the number of targets – and thus the sparsity level of the range profile – increases, Monte Carlo simulations proved that QIHT still provides high range estimation performances by promoting consistency with the quantized radar observations. As a proof of concept, we obtained similar range estimation performances from quantized observations of an actual radar in a controlled environment; hence showing that this QCS radar framework could apply in radar applications with limited bit-rate, *e.g.*, for radar signal reception with cheap ADCs. Future work will address the interplay between the dither and the background noise, with a practical realization of the proposed highly quantized and dithered architecture. On top of the noise, the other non-idealities introduced at the end of Chapter 2 should also be added to the study. For example, the dithering process heavily depends on the dynamic range of the received signal, which can be modified by the RF leakage.

Chapter 5

Range and Angle of Arrival Estimation

WE present a novel scheme allowing for 2D target localization using highly quantized 1-bit measurements from a *Frequency Modulated Continuous Wave* (FMCW) radar with two receiving antennas. Quantization of radar signals introduces localization artifacts; we remove this limitation by inserting a *dithering* on the unquantized observations. We then adapt the projected back projection algorithm to estimate both the range and angle of targets from the dithered quantized radar observations, with provably decaying reconstruction error when the number of observations increases. Simulations are performed to highlight the accuracy of the dithered scheme in noiseless conditions when compared to the non-dithered and full 32-bit resolution under severe bit-rate reduction. Measurements are performed using a radar sensor to demonstrate the effectiveness and performances of the proposed quantized dithered scheme in real conditions. Finally, to further reduce the hardware requirements and bit rate, we study the effect of dropping one of the baseband IQ channels from each receiving antenna. To that end, the structure of the received signals is exploited to recover the positions of multiple targets. Simulations are performed to highlight the accuracy and limitations of the proposed scheme under severe bit-rate reduction.

5.1 Introduction

Compressive sensing aims at compressively and non-adaptively sampling structured signals, *e.g.*, sparse or compressible signals in an appropriate basis, by correlating them with a few random patterns, *i.e.*, much less numerous than the ambient signal dimension [CRT06b]. The compressively observed signal is then estimated from non-linear algorithms such as basis pursuit denoise (BPDN) [CT05], iterative hard thresholding (IHT) [BD09], or CoSaMP [NT09].

In radar processing, CS offers the potential to simplify the acquisition process [BS07] or to use super resolution algorithms to solve ambiguous estimation problems [HS09]. However, the underlying assumption of such schemes is the availability of high resolution radar signals, requiring high bit-rate data transmission to a processing unit.

In this chapter, we aim to break this assumption and to further explore the reconstruction of the target scene on the basis of radar signals acquired under harsh bit-rate acquisition process, *i.e.*, a regime where classic estimation methods fail (*e.g.*, Maximum Likelihood [Kay93]). Bit-rate reduction in radar applications indeed opens new study directions, *e.g.*, through the use of 1-bit comparators to design cost-efficient acquisition hardware, or the use of several radar sensors run in parallel with fixed data-rate, as in *Internet of Things* (IOT) applications relying on a massive collection of sensors. Moreover, this loss of resolution can be counteracted by increasing the number of observations, provided that new algorithms be designed for this context.

We propose to reconstruct the target scene in the extreme 1-bit measurement regime, in a similar way to only recording the sign of each sample [BB08; Jac+13; PV13]. Comparing with the existing literature on 1-bit quantization of "IQ" signals for different radar applications [DZ15; Li+16], our main contributions lie in the following aspects. First, we show that estimating the 2D-localization of multiple targets observed from a radar system with two antennas under the harsh bit-rate requirement is feasible. This problem amounts to estimating a sparse signal, whose support and phases encode the target ranges and angles, from a quantized CS (QCS) model. In particular, we explore the estimation in an extreme bit-rate scenario where

every measurement takes a single bit achieved by a uniform scalar quantization combined with a random dithering vector [DJR19; XJ19; XSJ18].

Second, we provide theoretical guarantees on the estimation error of multiple targets localization using the *projected back projection* (PBP) algorithm [Bar+17; XJ19; XSJ18]. This is achieved by promoting in PBP a joint support between the range profiles observed by the two antennas. In particular, we show that the estimation error decays when the number of quantized observations increases. We further reveal, through Monte Carlo simulations, a certain trade-off between the number of measurements and the total bit-rate by comparing the performances of PBP under multiple scenarios involving one or two targets and different measurement numbers and resolutions. The importance of the dithering process is also stressed by the existence of strong artefacts in the 2D-localization of targets when this dithering is not added. We demonstrate our method in real experiments, locating corner reflectors in an anechoic chamber. In this context, we show that the random dithering still improves the localization of targets provided this dithering is adapted to the signal noise. Finally, by leveraging the structure between received signals from different antennas, we show that one can further simplify the acquisition process by only recording half of the signals outputted by the coherent demodulation, *i.e.*, by *Dropping Channels*. We show that this restriction does not affect the maximum recoverable range but rather only renders pairs of ranges possibly ambiguous in exchange of cutting in half the amount of data required for the estimation.

The rest of this chapter is structured as follows. The radar signal model presented in Chapter 2 and Chapter 4 are extended to the 2D setting in Sec. 5.2. The quantized radar observation model, the adaptation of the PBP algorithm to the 2D-localization of multiple targets and the theoretical analysis of its reconstruction error are provided in Sec. 5.3 and Sec. 5.4. In Sec. 5.5, the proposed scheme is tested under different scenarios using Monte Carlo simulations. Finally, we report the use of our framework in a real experiment in Sec. 5.6 and introduce the Channel Dropping acquisition in Sec. 5.7 before concluding.

5.2 Radar System Model

In this section, we show how the 2D target localization information is *linearly* encoded in the signals recorded by a radar system involving two antennas illustrated in Fig. 5.1. To that end, we extend the model of the received signal $r(t)$ in (2.8) introduced in Chapter 2 for one transmit and one receive antennas to the 2 receive antennas configurations.

Let us first consider one static target located at range $R > 0$ and angle $\theta \in [-\pi, \pi]$ from a receiving linear array comprised of two receiving antennas $\mathcal{Y}_1$ and $\mathcal{Y}_2$, located in $(0, 0)$ and $(0, d)$, respectively (see Fig. 5.1).

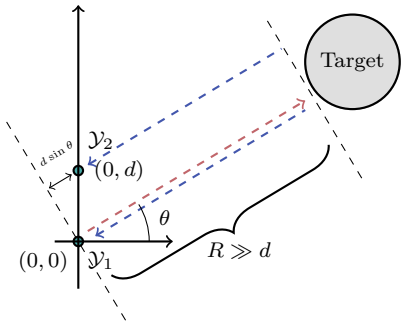


Figure 5.1: Illustration of the two antennas radar system with an array of receiving antennas.

The signal received on $\mathcal{Y}_p$ ($p \in \{1, 2\}$) is:

$$y_p(t) = A\, s(t - \tau_p),$$

where A is the complex received amplitude coefficient that depends on several parameters such as the range R and the *Radar Cross-Section* (RCS) and $s(t)$ is the transmitted signal defined in (2.2). Under the far-field approximation, the delay $\tau_p = \mathsf{c}^{-1}(2R + p\, d \sin\theta)$ is the round-trip time between the transmitting antenna in $(0, 0)$, the target and $\mathcal{Y}_p$ ($p \in \{1, 2\}$), with c the speed of light.

After coherent demodulation of the FMCW radar signals as in Section 2.3, the acquisition model links the sampling time with the frequency that is being transmitted. The sampled signal can be seen as a measurement of the phase-shift at time t of the transmitted frequency $f_c(t)$ depending on

the target position. For a regular sampling at rate N/T (similarly to Chapter 4), the sampled frequencies are $f_i := f_c(i\frac{T}{N})$, $1 \leq i \leq N$, so that, at the i^{th} frequency, $\mathcal{Y}_p$ receives the signal:

$$Y_{ip} = x\, e^{-\mathrm{i}\, 2\pi f_i \tau_p} \approx x\, e^{-\mathrm{i}\, 2\pi f_i \frac{2R}{c}} e^{-\mathrm{i}\, 2\pi f_0 \frac{pd \sin\theta}{c}}, \tag{5.1}$$

where x is the received amplitude after the coherent demodulation. The approximation in (5.1) is reasonable for K-band radars where $B \ll f_0$, *i.e.*, $B = 250\text{MHz}$ and $f_0 = 24\text{GHz}$ respectively.

Comparing (5.1) for $\mathcal{Y}_1$ and $\mathcal{Y}_2$ shows that the angle of arrival acts as a complex gain on $\mathcal{Y}_2$. Furthermore, this chapter considers a multi-target scenario using a purely additive model. This means that all the targets are in a direct line of sight from the radar, *i.e.*, there is no multi-path. For a scene with K targets, recasting (5.1) into a linear matrix sensing model and taking advantages of the phase relation between $\mathcal{Y}_1$ and $\mathcal{Y}_2$, the sensed signals $\boldsymbol{Y} = \{Y_{ip}\}_{ip} \in \mathbb{C}^{m\times 2}$ are

$$\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2] = \boldsymbol{\Phi}\,[\boldsymbol{x}, \boldsymbol{G}\boldsymbol{x}], \tag{5.2}$$

where $\boldsymbol{x} \in \mathbb{C}^N$ encodes the range profile, *i.e.*, $x_n \neq 0$ if there exists a target at range R_n, $\|\boldsymbol{x}\|_0 := |\operatorname{supp} \boldsymbol{x}| \leq s \ll N$, $\boldsymbol{\Phi} = \{e^{-\mathrm{i}\,\frac{4\pi}{c} f_i R_n}\}_{in} \in \mathbb{C}^{m\times N}$ is the *range* measurement matrix, $\boldsymbol{G} = \operatorname{diag}(e^{\mathrm{i}\,\frac{2\pi}{c} f_0 d \sin\theta_1}, \cdots, e^{\mathrm{i}\,\frac{2\pi}{c} f_0 d \sin\theta_N})$ with $\operatorname{supp}(\boldsymbol{\theta}) = \operatorname{supp}(\boldsymbol{x})$, *i.e.*, $\boldsymbol{G}$ is the phase difference between the first and second receiving antennas. Therefore, the 2D-localization problem is tantamount to estimating the support T of $\boldsymbol{x}$ from the sensing model (5.2), hence extracting the target ranges according to the discretization $\{R_n\}$. Comparing the phases of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ on the index set T then allows to deduce the angles $\boldsymbol{\theta}_T$. Interestingly, in this process, only the target ranges are discretized, *i.e.*, the angles are estimated from continuous phase differences. This, however, comes at a cost as this simplified two-antenna model does not allow the recovery of multiple targets located on the same range R_n.

Note that, in the absence of quantization, inverting (5.2) can be solved using Maximum Likelihood [Kay93] if $m \geq N$, or using CS algorithms (*e.g.*, IHT [BD09] or CoSaMP [NT09]) if $m \leq N$.

5.3 Quantizing Radar observations

In this work, we propose to quantize the observations $\boldsymbol{Y}$ achieved in the digital beamforming model (5.2). Our quantization procedure relies on the 1-bit quantizer with dithering $\mathcal{Q}_\epsilon^+$ defined in Chapter 3. The resolution ϵ is according to the dynamic of the received signals *i.e.*, $\epsilon > 2\max(\|\boldsymbol{\Phi}\boldsymbol{x}_1\|_\infty, \|\boldsymbol{\Phi}\boldsymbol{x}_2\|_\infty)$.

Our global objective is thus to estimate the localization of targets, as encoded in the matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2) = (\boldsymbol{x}, \boldsymbol{G}\boldsymbol{x}) \in \mathbb{C}^{N\times 2}$, from the quantized observation model

$$\boldsymbol{Z} = \mathcal{Q}_\epsilon^+(\boldsymbol{Y}) := \big(\mathcal{Q}_\epsilon^+(\boldsymbol{\Phi}\boldsymbol{x}_1), \mathcal{Q}_\epsilon^+(\boldsymbol{\Phi}\boldsymbol{x}_2)\big). \tag{5.3}$$

In $\mathcal{Q}_\epsilon^+$, a uniform random *dithering* $\boldsymbol{\xi} \in \mathbb{C}^m$, *i.e.*, $\xi_i \sim_{\text{i.i.d.}} \mathcal{U}_\epsilon^{\mathbb{C}}$ for all $i \in [m]$, is added to the quantizer input. The signals received on the different antennas are dithered with different random dithering. For real sensing models, such a dithering attenuates the impact of the quantizer on the estimation of sparse/compressible signals in quantized CS [Bou12; JC17; XJ19], the previous chapter also showed that dithering is necessary in ranging applications in order to be able to recover all sparse scenes. As will be clearer below, $\boldsymbol{\xi}$ also enables accurate estimation of $\boldsymbol{X}$, thus not only the range but the angle of arrival as well.

As written in (5.3), we can identify a low-complexity model for $\boldsymbol{X}$ in the case where only s targets, with distinct ranges, are observed. We quickly see that

$$\boldsymbol{X} \in \Theta^s := \bigcup_{T\subset[N],|T|\leq s} \Theta_T,$$

with $\Theta_T := \{\boldsymbol{U} \in \mathbb{C}^{N\times 2} : \operatorname{supp}(\boldsymbol{U}_{:,1}) = \operatorname{supp}(\boldsymbol{U}_{:,2}) \subset T\}$, which is a union of $\binom{N}{s}$ s-dimensional subspaces. Note that, according to (5.2), since $|G_{jj}| = 1$ we could further impose $|X_{j1}| = |X_{j2}|$ for all $j \in [N]$. However, this leads to an hardly integrable non-convex constraint on the domain of $\boldsymbol{X}$.

The next sections assess the quality of the reconstruction that can be obtained using these quantized measurements. To that end, we highlight the link between this 1-bit quantization $\mathcal{Q}_\epsilon^+$ and the mid-rise multibit ac-

quisition of resolution ϵ. Because of the careful scaling of the resolution ϵ, these two acquisitions are identical ($\mathcal{Q}_\epsilon = \mathcal{Q}^1_\epsilon$) and thus, allow us to use the framework developed in [XJ19].

5.4 2D Target Localization in Quantized Radar

Despite the quantization, the sensing model (5.3) still enables target localization. We adopt here a simple method, the projected back projection (PBP) proposed in [XJ19], for which the estimation error provably decays when the number of observations m increases. The PBP estimate is defined from

$$\hat{\boldsymbol{X}} = \mathcal{P}_{\Theta^s}(\tfrac{1}{m}\boldsymbol{\Phi}^H \boldsymbol{Z}), \tag{5.4}$$

with the projector $\mathcal{P}_{\Theta^s}(\boldsymbol{U}) := \arg\min_{\boldsymbol{V}\in\Theta^s} \|\boldsymbol{U}-\boldsymbol{V}\|_F^2$. In words, the estimate $\hat{\boldsymbol{X}}$ is achieved by first *back projecting* $\boldsymbol{Z}$ in the signal domain thanks to the adjoint sensing $\boldsymbol{\Phi}^H$, and then taking the closest point in Θ^s to $\frac{1}{m}\boldsymbol{\Phi}^H\boldsymbol{Z}$ from the projector $\mathcal{P}_{\Theta^s}$.

Interestingly, for any $\boldsymbol{U} \in \mathbb{C}^{N\times 2}$, $\boldsymbol{V} = \mathcal{P}_{\Theta^s}(\boldsymbol{U})$ is easily computed. Denoting by $\mathsf{H}_s(\boldsymbol{v})$ the hard-thresholding operator setting all but the s largest components (in magnitude) of $\boldsymbol{v} \in \mathbb{C}^N$ to zero, we first form $T = \operatorname{supp}(\mathsf{H}_s(\boldsymbol{u}))$ with $\boldsymbol{u} = \mathcal{S}(\boldsymbol{U}) \in \mathbb{R}^N_+$ and $\mathcal{S}(\boldsymbol{U})_n = (|U_{n,1}|^2 + |U_{n,2}|^2)^{1/2}$ for all $n \in [N]$, and then, for $p \in \{1,2\}$, V_{np} equals to U_{np} if $n \in T$, and to 0 otherwise. This provides clearly $\|\mathcal{P}_{\Theta^s}(\boldsymbol{U}) - \boldsymbol{U}\|_F^2 = \|\mathsf{H}_s(\mathcal{S}(\boldsymbol{U})) - \mathcal{S}(\boldsymbol{U})\|_2^2 \le \|\tilde{\boldsymbol{u}} - \mathcal{S}(\boldsymbol{U})\|_2^2$ for any $\tilde{\boldsymbol{u}} \in \mathbb{R}^n_+$ such that $|\operatorname{supp}(\tilde{\boldsymbol{u}})| \le K$. Since for any $\boldsymbol{U}' \in \Theta^s$, $\tilde{\boldsymbol{u}} := \mathcal{S}(\boldsymbol{U}') \in \mathbb{R}^n_+$, we thus have $\|\mathcal{P}_{\Theta^s}(\boldsymbol{U}) - \boldsymbol{U}\|_F^2 \le \|\mathcal{S}(\boldsymbol{U}') - \mathcal{S}(\boldsymbol{U})\|_2^2 = \|\boldsymbol{U}' - \boldsymbol{U}\|_F^2$ as required from the definition of $\mathcal{P}_{\Theta^s}$.

Given the estimate $\hat{\boldsymbol{X}}$ in (5.4), the range profile $\hat{T}$ is simply obtained as $\hat{T} = \operatorname{supp}\big(\mathcal{S}(\hat{\boldsymbol{X}})\big)$, so that targets are localized in the polar coordinates (R_n, θ_n) for all $n \in \hat{T}$ with $\theta_n = \arcsin\big(\frac{c}{2\pi f_0 d}\angle(\hat{x}_2[n]^*\hat{x}_1[n])\big)$.

We now establish how the estimation error $\|\hat{\boldsymbol{X}} - \boldsymbol{X}\|_F$ of (5.4) can be bounded with high probability. This is important to ensure the quality of the estimated target coordinates. To this end, given $0 < \epsilon < 1$, we first assume that our radar sensing matrix $\frac{1}{\sqrt{m}}\boldsymbol{\Phi}$ respects the restricted isometry property over the set $\bar{\Sigma}^N_s := \{\boldsymbol{u} \in \mathbb{C}^N : |\operatorname{supp}\boldsymbol{u}| \le s\}$ of complex s-sparse signals, in short $\frac{1}{\sqrt{m}}\boldsymbol{\Phi} \in \mathrm{RIP}(\bar{\Sigma}^N_s, \epsilon)$.

Many random matrix constructions have been proved to respect the

RIP with high probability (w.h.p.[1]) [Bar+08; FR13; CRT06b]. Given the discrete Fourier matrix $\boldsymbol{F} \in \mathbb{C}^{N\times N}$, the selection matrix $\boldsymbol{R}_\Omega$ such that $\boldsymbol{R}_\Omega \boldsymbol{u} = \boldsymbol{u}_\Omega$, and provided $m \gtrsim \delta^{-2} s\,(\ln s)^2 \ln N$, if $\Omega \subset [N]$ has cardinality m and is picked uniformly at random among the $\binom{N}{m}$ m-length subsets of $[N]$, then $\boldsymbol{\Phi} = \sqrt{N}\boldsymbol{R}_\Omega \boldsymbol{F}$ respects w.h.p. the RIP$(\bar{\Sigma}_s^N, \delta)$ [FR13; Rau10]. Therefore, up to a random sub-sampling of the m frequencies $\{f_i\}$, the radar sensing matrix $\boldsymbol{\Phi}$ follows a similar construction.

Second, given $\nu > 0$, we assume that $\mathcal{A}_b^+ := \mathcal{Q}_\epsilon^b(\boldsymbol{\Phi}\cdot + \boldsymbol{\xi})$ satisfies the (complex) limited projection distortion over $\bar{\Sigma}_s^N$, or $\mathcal{A}_b^+ \in \text{LPD}(\bar{\Sigma}_s^N, \boldsymbol{\Phi}, \nu)$, *i.e.*,

$$\tfrac{1}{m}|\langle \mathcal{A}_b^+(\boldsymbol{w}), \boldsymbol{\Phi}\boldsymbol{v}\rangle - \langle \boldsymbol{\Phi}\boldsymbol{w}, \boldsymbol{\Phi}\boldsymbol{v}\rangle| \le \nu,\ \forall \boldsymbol{w}, \boldsymbol{v} \in \bar{\Sigma}^N \cap \mathbb{B}^N. \tag{5.5}$$

Thanks to these two conditions, we get the following guarantee on $\hat{\boldsymbol{X}}$.

Lemma 5.1. *Given $\delta, \nu > 0$, if $\frac{1}{\sqrt{m}}\boldsymbol{\Phi} \in \text{RIP}(\bar{\Sigma}_{2s}^N, \delta)$ and $\mathcal{A}_b^+ \in \text{LPD}(\bar{\Sigma}_{2s}^N, \boldsymbol{\Phi}, \nu)$, then, for all $\boldsymbol{X} \in \Theta_s$ the PBP estimate (5.4) satisfies $\|\hat{\boldsymbol{X}} - \boldsymbol{X}\|_F \le 2(\delta + 2\nu)$.*

The next proposition determines when $\mathcal{A}_b^+$ respects the LPD, as required by Lemma. 5.1.

Lemma 5.2. *Given $\delta > 0$, if $\frac{1}{\sqrt{m}}\boldsymbol{\Phi} \in \text{RIP}(\bar{\Sigma}_{2s}^N, \delta)$ and if $m \ge C\delta^{-2} s \ln(\frac{N}{s}) \ln(1 + \frac{c}{\delta^3})$, then, w.h.p., $\mathcal{A}_b^+ \in \text{LPD}(\bar{\Sigma}_s^N, \boldsymbol{\Phi}, 4\delta(1+\epsilon))$.*

In other words, inverting the role of m and ϵ in the requirement of Prop. 5.2 and assuming $\frac{1}{\sqrt{m}}\boldsymbol{\Phi}$ is RIP, up to some log factors, Prop. 5.1 shows that, w.h.p., the estimation error of PBP decays like $\|\hat{\boldsymbol{X}} - \boldsymbol{X}\|_F = O\big((1+\epsilon)\sqrt{s/m}\big)$ when m increases.

[1] "w.h.p." means with probability exceeding $1 - Ce^{-c\epsilon^2 m}$ for $C, c > 0$.

5.5 Numerical Results

In this section, Monte Carlo simulations are performed for different sparsity levels to assess the accuracy of the proposed scheme for a variety of targets' positions.

5.5.1 Parameters and metrics

We simulate the working mode of a noiseless K-Band radar, *i.e.*, giving $f_0 = 24\text{GHz}$ and a bandwidth of $B = 250\text{MHz}$. The spacing d between the two antennas is defined as half a wavelength, *i.e.*, $d = \frac{c}{2f_0}$, allowing for angular estimation in $[-\frac{\pi}{2}, \frac{\pi}{2}]$. In all our simulations, we set the number of ranges N to 256, giving a range limit of $R_{\max} = 153.6\text{m}$ and a range resolution of 0.6m. We test $320\,000 \times s$ Monte Carlo runs, where s is the considered sparsity, the targets' localization are picked uniformly at random in a 40×40 discretized polar domain $(R, \theta) \in [0, R_{\max}] \times [-\pi/2, \pi/2]$. In this domain, all targets receive uniformly random phases in $[0, 2\pi]$, the strongest target being set to a unit amplitude and the weaker ones having uniformly distributed amplitudes in $[0, 1]$. In order to focus on bit-rate reduction in radar processing, a total budget of 512 bits per channel is fixed for each acquisition with m measurements, *i.e.*, giving $m = 512$ measurements for 1-bit measurement quantization (*i.e.*, 2 complete FMCW saw-tooths), or $m = 16$ for 32-bit measurements. Our regime thus leads to a bit-rate reduction of 93.75% compared to a full acquisition with $m = 256$ for 32-bit measurements. The quality of the position estimation is simply measured as $\min_k |Re^{\mathrm{i}\,\theta} - \hat{R}_k e^{\mathrm{i}\,\hat{\theta}_k}|$, *i.e.*, the distance between the true target location and the closest estimated targets in $(\hat{\boldsymbol{R}}, \hat{\boldsymbol{\theta}})$. This quality measure is then averaged over runs which have the same position (R, θ). These results are reported in a 2D polar graph (Fig. 5.2, Fig. 5.3, Fig. 5.4).

5.5.2 Simulations for a Single Target Scenario

In this first simulation, we test the 2D-localization of a single target (*i.e.*, $\boldsymbol{X} \in \Theta^1$) for the dithered and non-dithered schemes. In Fig. 5.2b, the non-dithered scheme exhibits systematic artifacts in the estimation quality. Indeed, in the context of radar localization of one target, ambiguities

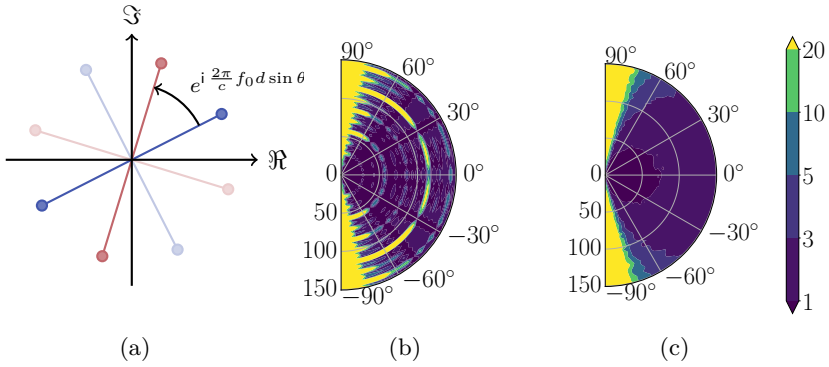


Figure 5.2: (a) Example of a possible 2D target localization ambiguity. (b) and (c), positions error in meters for Monte Carlo simulations with one target and $M = 512$, for 1-bit non-dithered and 1-bit dithered quantization scheme, respectively.

appear when full resolution signals from different receiving antennas once quantized are the same. This is similar to the limitations encountered by the deterministic quantization that were highlighted in the previous chapter (see Section 4.3). One obvious possibility is when the angle of arrival θ is so small that $\boldsymbol{G} \approx \mathbf{Id}$ which means that $\mathcal{Q}_\epsilon(\boldsymbol{\Phi x}) = \mathcal{Q}_\epsilon(\boldsymbol{\Phi G x})$. But this ambiguity can only appear for small angles of arrival and does not depend on the range. The artefacts present in Fig. 5.2b are highly structured and depend on the range. Another possibility appears for targets with large angles of arrival at certain ranges, as seen in the artifact pattern in Fig. 5.2b. Certain ranges induce a strong repetition between quantized measurements, as depicted for illustration in Fig. 5.2a for a range of $\frac{1}{4}R_{\max}$. The measurements in **blue** correspond to the signal coming from the first antenna and the one in **red** is the phase-shifted signal received on the second antenna. Without dithering, these signals lie in the same quadrants, even for substantially large phase-shift, and thus large θ. The quantized signals being identical, the estimated angle is 0° regardless of the actual angle. This ambiguity is the equivalent of the one presented in Chapter 4 consider a 1D problem with more than one target but here the 2D localization problem is considered and ambiguities appear with only one target.

The dithered scheme in Fig. 5.2c exhibits good performances on a wide range of positions and the effect of the dithering is clearly visible by the absence of artefacts. The drop of performances in Fig. 5.2b and Fig. 5.2c

at $\pm 72°$ degrees is related to the sensitivity of the arcsin function.

5.5.3 Simulations for the 2 Targets Scenario

Fig. 5.3 shows the performances of the schemes for 2 targets. When compared to the non-dithered (Fig. 5.3a) or full resolution (Fig. 5.3c), the dithered strategy in Fig. 5.3b surpasses the others constrained to the same bit rate. Comparing Fig. 5.2c and Fig. 5.3b, a drop of performances can

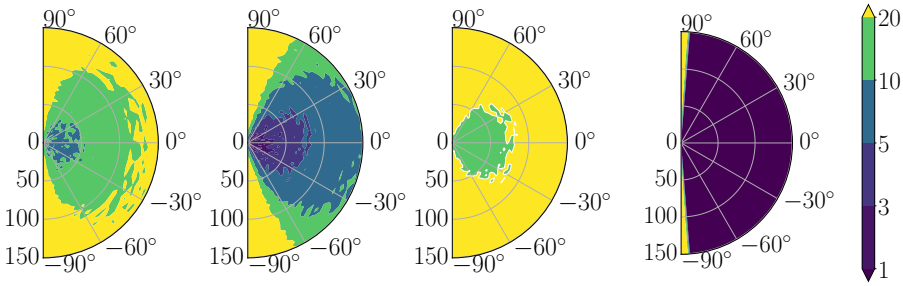


Figure 5.3: Positions error in meters for Monte Carlo simulations with two targets; (a) $m = 512$, 1-bit non-dithered quantization; (b) $m = 512$, 1-bit dithered quantization; (c) $m = 16$, 32-bit non-dithered; (d) $m = 256$, 32-bit full measurements.

be seen from the increase in sparsity. This is consistent with the results in Sec. 5.3, where we showed that the bound on the error of PBP grows as the sparsity increases. Moreover, in the absence of dithering in our quantized radar scheme, extremely sparse signals can lead to ambiguous estimations as was showcased in the previous Chapter (see Chap. 4). In Fig. 5.4, the strongest and weakest estimated targets are separated to study their respective accuracy. For the non-dithered scheme in Fig. 5.4a the strongest target still exhibits artifacts whereas the weakest (Fig. 5.4b) is consistently wrongly estimated. The dithering reduces partly these situations and offers better performances for both targets in Fig. 5.4c and Fig. 5.4d. While the accuracy of the second target for the dithered scheme is impressive when compared to the non-dithered one, it is far from what can be achieved in Fig. 5.3d for the full measurements and resolution approach. This work is one of the first venture into radar localization using 1-bit dithered scheme that is, furthermore, constrained to a specifically low bit-rate. In the fu-

ture, better reconstruction qualities could be obtained by replacing PBP with other algorithms explicitly using the dithering to reach consistent signal estimates [DJR19; Bar+17]. The reconstruction could be enhanced using the quantized version of IHT [BD09], as was done in Chapter 4.

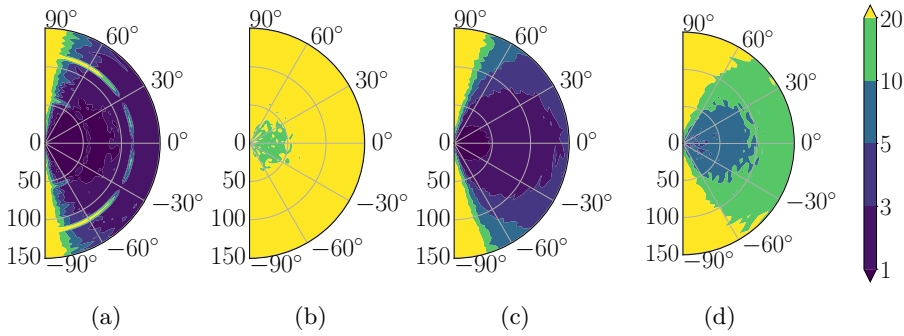


Figure 5.4: Positions error in meters for Monte Carlo simulations with two targets, $m = 512$ and 1-bit quantization; in (a) and (b), strongest and weakest target for the non-dithered scheme, respectively; in (c) and (d), strongest and weakest target for the dithered scheme, respectively.

5.6 Experimental Validation

The study of the proposed scheme is now extended to real measurements to test the model and reconstruction algorithm against noise and non-idealities from the environment, the targets and the radar (see Sec. 2.3).

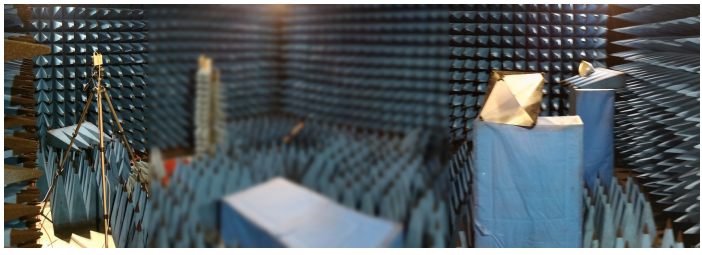
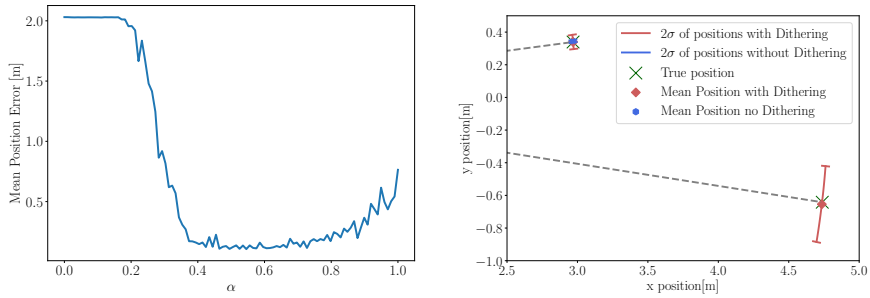


Figure 5.5: Experiment set-up with a FMCW radar on the left and two corner reflectors on the right.

The measurements are performed in an anechoic chamber where two targets are located in front of a commercial radar product [RFB] at different



(a) (b)

Figure 5.6: Reconstruction using real measurements, (a) mean positions error for different levels of dithering; (b) reconstructions achieved with weighted dithering.

ranges and angles (see Fig.5.5). The radar parameters, *e.g.*, B and N, mirror the ones used in the simulations. The proposed scheme and the developed theory only considers the noiseless case. Practical measurements with real radar sensors are, however, inherently corrupted by noise. To this end, the reconstruction is studied with different levels of added dithering to assess the impact of the already present noise on the quantization. Fig. 5.6a shows the mean position errors for the two targets for a weighted dithering $\tilde{\boldsymbol{\xi}} = \alpha\boldsymbol{\xi}$, with $\alpha \in [0, 1]$, where $\boldsymbol{\xi}$ is the dithering defined in Sec. 5.3. Fig. 5.6a shows that a certain amount of dithering is required to achieve good performance but also that adding a full dithering (*i.e.*, $\alpha = 1$) is not the *by default* optimum. Fig. 5.6b shows the reconstruction achieved using the optimal scaling $\alpha = 0.55$ of the dithering versus the absence of dithering ($\alpha = 0$). The radar is located in $(0, 0)$. The variations in the estimations for 106 consecutive measurements are represented by the two sigma span around the mean estimated positions. As already hinted in Fig. 5.6a, the non-dithered scheme exhibits poor performances as it is not able to resolve the second target at the 4.8m range. The 1-bit non-dithered quantization has effectively removed the second target from the signal. It is interesting to see that the effect obtained in Fig. 4.5 in the previous chapter, with an actual radar and target simulators are replicated using here a fully practical setting. Adding the weighted dithering allows the recovery of the two targets consistently but at a price in the variance of the closest one. This result shows nonetheless a promising gain of the use of dithering on real applications where noise is

encountered.

5.7 Channel Dropping Model

In this section, we expand on the result presented in the first part of this chapter by studying a modified acquisition setting of an FMCW radar with two receiving antennas.

5.7.1 Modified Signal and Quantization Model

The coherent demodulation of the signals received by the antennas produces complex signals that are each transmitted on 2 channels representing the real and the imaginary parts as depicted in Fig. 2.5. Common radar acquisition scheme requires the sampling of both of these channels, resulting in $2m$ measurements per antenna. We propose to acquire only half of the channels as follows:

$$\tilde{\boldsymbol{y}}_1 = \mathrm{Re}\{\boldsymbol{y}_1\}, \quad \tilde{\boldsymbol{y}}_2 = \mathsf{i}\,\mathrm{Im}\{\boldsymbol{y}_2\} \tag{5.6}$$

This simplification, however, comes at a cost. As explained in [Feu+18a], range estimation is equivalent to estimating the support of the spectrum of the received signals. $\boldsymbol{y}_1$ and $-\mathsf{i}\,\boldsymbol{y}_2$, being purely real signals, their spectrums are by definition symmetric, doubling the sparsity in an ambiguous way. In the spectrum of these *incomplete* measurements, a target at a range R will appear at R and $R_{max} - R$ with a complex amplitude of α and α^* respectively. To partly solve this problem, we observe that for one target at (R_k, θ) :

$$\begin{aligned} \hat{\boldsymbol{y}} &= \tilde{\boldsymbol{y}}_1 + \tilde{\boldsymbol{y}}_2 \\ &= (1 + G_{kk})\, e^{-\mathsf{i}\, 2\pi f_m \frac{2R_k}{\mathsf{c}}} + (1 - G_{kk}^*)\, e^{\mathsf{i}\, 2\pi f_m \frac{2R_k}{\mathsf{c}}} \end{aligned} \tag{5.7}$$

Eq.(5.7) shows, provided G tends to 1 (*i.e.*, for small angles) that the unambiguous support of the complex signal can be approximated (*i.e.*, $\mathrm{supp}(\boldsymbol{\Phi}^H \hat{\boldsymbol{y}}) \approx \mathrm{supp}(\boldsymbol{x}_1)$). Indeed, if $G_{kk} \approx 1$ then one can expect that $|1 + G_{kk}| \geq |1 - G_{kk}^*|$, which a hard-thresholding operator can recover successfully. The next section introduces a variant of the hard-thresholding operator H_s that can deal with this symmetry for $s \geq 2$ and its impact on

the performances.

It is important to stress that, given the Fourier nature of (5.2) some ranges will remain ambiguous *e.g.*, for ranges $R = 0$ and $R = R_{max}/2$ and that this limitation affects the signal regardless of the quantization.

In this work, we propose to quantize the observations $\boldsymbol{Y}$ achieved in the digital beamforming model (5.6). Our global objective is thus to estimate the localization of targets, as encoded in the matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2) = (\boldsymbol{x}, \boldsymbol{G}\boldsymbol{x}) \in \mathbb{C}^{N\times 2}$, from the quantized observation model

$$\boldsymbol{Z} = \tilde{\mathcal{Q}}_\epsilon^+(\boldsymbol{\Phi}\boldsymbol{X}) := \big(\mathrm{Re}\{\mathcal{Q}_\epsilon^+(\boldsymbol{\Phi}\boldsymbol{x}_1)\}, \mathrm{i}\,\mathrm{Im}\{\mathcal{Q}_\epsilon^+(\boldsymbol{\Phi}\boldsymbol{x}_2)\}\big). \tag{5.8}$$

5.7.2 Modified 2D Target Localization in Quantized Radar

Given the altered structure of the signal caused by the *Dropping Channel* model and the ability of the estimate in (5.7) to recover the ranges of targets, one needs to adapt the reconstruction algorithm. We adopt here a simple method, which is an adaptation of PBP proposed in [XJ19] and that was used successfully in this chapter to recover the 2D position of targets. The estimate is, first, defined from the back-projection:

$$\hat{\boldsymbol{X}} = \tfrac{1}{m}\boldsymbol{\Phi}^H\boldsymbol{Z}. \tag{5.9}$$

Next, to recover the support $\hat{T}$ from $\hat{\boldsymbol{X}}$ we rely on the approximation defined in (5.7) to partly cancel the symmetrical shape of $\hat{\boldsymbol{x}}_1$ and $\hat{\boldsymbol{x}}_2$:

$$\hat{T} = \mathrm{supp}\left(\mathsf{H}_s^{\mathrm{Sym}}(\hat{\boldsymbol{x}}_1 + \hat{\boldsymbol{x}}_2)\right) \tag{5.10}$$

where $\mathsf{H}_s^{\mathrm{Sym}}(\cdot)$ is the hard-thresholding operator which takes the s biggest elements excluding the weakest symmetrical elements. An example of the application of this operator is provided in Fig. 5.7. The symmetric pairs of ambiguous targets are centred around $\frac{1}{2}R_{max}$ and the hard-thresholding operator $\mathsf{H}_s^{\mathrm{Sym}}$ selects the biggest (in amplitude) of the pairs and ignores the other. This allows the detection of targets on the whole range $[0, R_{max}]$ but with the added restriction that targets lying at range symmetrical from $\frac{1}{2}R_{max}$ will be ambiguous.

It is interesting to note that some radar modules only samples the real

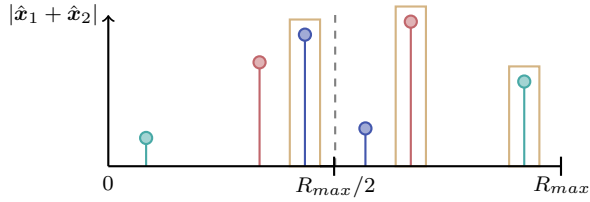


Figure 5.7: Example of the selection operated by $\mathsf{H}_s^{\mathrm{Sym}}$ for a vector $\hat{x}_1 + \hat{x}_2$ estimated from the measurements of a 3-sparse vector; each pair of ambiguous peak are in different colours and the selected peaks are highlighted in yellow.

parts of the demodulated signals, which simplifies the demodulation procedure presented in Fig. 2.5. This impacts the estimation in a different way. Because of the symmetry introduced by the strictly real measurements, the unambiguous maximum range is $\frac{R_{max}}{2}$ while the proposed model keeps the maximum range.

The targets are localized in the polar coordinates (R_n, θ_n) for all $n \in \hat{T}$, with $\theta_n = \arcsin\left(\frac{c}{2\pi f_0 d} \angle(\hat{x}_2^*[n]\hat{x}_1[n])\right)$.

5.7.3 Simulations Results

The simulations setting and parameters are identical to the one used in Sec. 5.5 and in [Feu+18a]. Fig. 5.8 shows the performances of the proposed

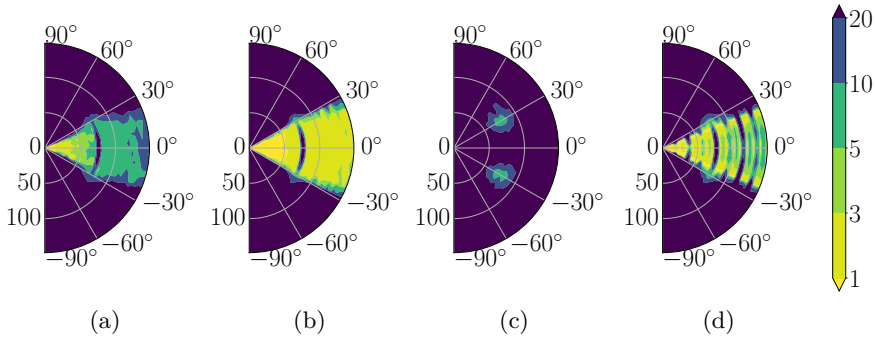


Figure 5.8: Positions error in meters for Monte Carlo simulations with one target; (a) 1-bit dithered with $\frac{m}{N} = 20\%$; (b) 1-bit dithered with $\frac{m}{N} = 200\%$; (c) 1-bit dithered with $\frac{m}{N} = 200\%$ using PBP in [Feu+18a]; (d) 1-bit non-dithered with $\frac{m}{N} = 200\%$, respectively.

scheme for different configurations for $s = 1$. Fig. 7.12b and Fig. 7.12a

show that the localization error decreases as m increases. Furthermore the bit-rate reduction compared to classic sampling scheme is of 99.69% and 96.87% respectively. Fig. 5.8c is the performance obtained when the scheme in [Feu+18a] is applied after channel dropping. The hard thresholding operator introduced in Eq.(5.10) is shown to be the key to recover from the omissions of channels. The artifacts showcased in [Feu+18a] and in Sec. 5.5 resulting from the 1-bit non dithered quantization are still present in Fig. 5.8d. The maximum angle that can be recovered is also reduced from [Feu+18a] which is consistent with the approximation (5.7). The sparsity

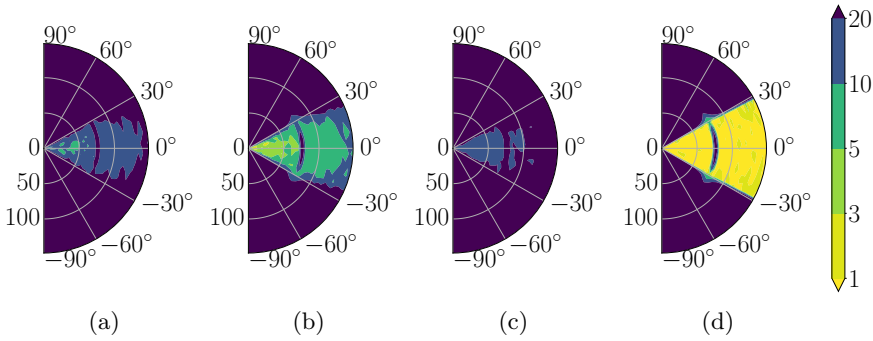


Figure 5.9: Positions error in meters for Monte Carlo simulations with two targets; (a) 1-bit non-dithered with $\frac{m}{N} = 200\%$; (b) 1-bit dithered with $\frac{m}{N} = 200\%$; (c) 32 bit non-dithered with $\frac{m}{N} = 6.25\%$; (d) 32 bit non-dithered with $\frac{m}{N} = 100\%$, respectively.

is increased to $s = 2$ in Fig. 5.9. Similarly to what has been observed in [Feu+18a], the dithered scheme in Fig. 5.9b clearly outperforms the non-dithered one in Fig. 5.9a. Fig. 5.9c shows the performances of the scheme when constrained to the same bit rate, thus with a smaller m, with a classic full acquisition scheme. The 1-bit dithered scheme represents an interesting trade off between the bit resolution and the transmitted bit-rate when compared to Fig. 5.9d. Finally the ambiguity in $\frac{R_{max}}{2}$ is present, as expected in every presented schemes.

5.8 Discussion

In this chapter, we have studied the 2D-localization of multiple targets configurations by using two receiving antennas combined with 1-bit radar

quantization, which resulted into the QCS model (5.8). We proved that the PBP algorithm for the 2D-localization of targets achieves a bounded reconstruction error decaying as the number of measurements increases. This decaying reconstruction error of the PBP algorithm was further verified using the Monte Carlo simulations and real radar measurements. In particular, the real radar measurements experiments with the radar sensor shed light on the interaction between the system noise and the uniform dithering. Furthermore, we showed how some deterministic artifacts vanish when a random dithering vector is added in the quantization process. In the second part of this chapter we showed that on top of lowering the resolution of the measurements, the structure of the beam-forming model can be leveraged in order to only record half of the channels (corresponding to the real and imaginary domain) and we showed that the reconstruction only suffered from minor ambiguities. Future work regarding this work will study the extension of this scheme beyond the 2D estimation of target. Indeed, recording consecutive FMCW chirps can also be used to estimate the velocity of targets thanks to the Doppler Effect. This work also showcased the impact of the noise on the dithered quantization. An in-depth study of the interplay between this added uniform random variable and the already present Gaussian noise is one of the next steps towards using 1-bit radar in more complex applications.

5.9 Proofs

Lemma 5.1. *Given* $\delta, \nu > 0$, *if* $\frac{1}{\sqrt{m}}\boldsymbol{\Phi} \in \mathrm{RIP}(\bar{\Sigma}_{2s}^N, \delta)$ *and* $\mathcal{A}_b^+ \in \mathrm{LPD}(\bar{\Sigma}_{2s}^N, \boldsymbol{\Phi}, \nu)$, *then, for all* $\boldsymbol{X} \in \Theta_s$ *the PBP estimate* (5.4) *satisfies* $\|\hat{\boldsymbol{X}} - \boldsymbol{X}\|_F \leq 2(\delta + 2\nu)$.

Proof. If $\frac{1}{\sqrt{m}}\boldsymbol{\Phi} \in \mathrm{RIP}(\bar{\Sigma}_{2s}^N, \delta)$, then $\frac{1}{\sqrt{m}}\boldsymbol{\Phi} \in \mathrm{RIP}(\Theta^{2s}, \delta)$ with respect to the Frobenius norm since

$$\begin{aligned} |\tfrac{1}{m}\|\boldsymbol{\Phi}\boldsymbol{U}\|_F^2 - \|\boldsymbol{U}\|_F^2| &\leq |\tfrac{1}{m}\|\boldsymbol{\Phi}\boldsymbol{u}_1\|^2 - \|\boldsymbol{u}_1\|^2| + |\tfrac{1}{M}\|\boldsymbol{\Phi}\boldsymbol{u}_2\|^2 - \|\boldsymbol{u}_1\|^2| \\ &\leq \delta(\|\boldsymbol{u}_1\|^2 + \|\boldsymbol{u}_2\|^2) = \delta\|\boldsymbol{U}\|_F^2, \end{aligned}$$

for all $\boldsymbol{U} = (\boldsymbol{u}_1, \boldsymbol{u}_2) \in \Theta^{2s} \subset (\bar{\Sigma}^N_{2s}, \bar{\Sigma}^N_{2s})$. Moreover, extending the LPD (5.5) to matrices with the Frobenius scalar product, if $\mathcal{A}^+_b \in \mathrm{LPD}(\bar{\Sigma}^N_{2s}, \boldsymbol{\Phi}, \nu)$, then the matrix map $\mathcal{A}^+_b \in \mathrm{LPD}(\Theta^{2s}, \boldsymbol{\Phi}, 2\nu)$ since $\langle \mathcal{A}^+_b(\boldsymbol{W}), \boldsymbol{\Phi}\boldsymbol{V}\rangle_F = \langle \mathcal{A}^+_b(\boldsymbol{w}_1), \boldsymbol{\Phi}\boldsymbol{v}_1\rangle + \langle \mathcal{A}^+_b(\boldsymbol{w}_2), \boldsymbol{\Phi}\boldsymbol{v}_2\rangle$ for any $\boldsymbol{W} = (\boldsymbol{w}_1, \boldsymbol{w}_2), \boldsymbol{V} = (\boldsymbol{v}_1, \boldsymbol{v}_2) \in \Theta^s$, and similarly for $\langle \boldsymbol{\Phi}\boldsymbol{W}, \boldsymbol{\Phi}\boldsymbol{V}\rangle_F$. The rest of the proof is a quick extension of [XJ19, Thm. 4.1] to complex $N \times 2$ matrices belonging to the union of low-dimensional subspaces Θ^s. □

Lemma 5.2. *Given $\delta > 0$, if $\frac{1}{\sqrt{m}}\boldsymbol{\Phi} \in \mathrm{RIP}(\bar{\Sigma}^N_{2s}, \delta)$ and if $m \geq C\delta^{-2} s \ln(\frac{N}{s}) \ln(1 + \frac{c}{\delta^3})$, then, w.h.p., $\mathcal{A}^+_b \in \mathrm{LPD}(\bar{\Sigma}^N_s, \boldsymbol{\Phi}, 4\delta(1+\epsilon))$.*

Proof. Extending the LPD to real mappings, we first note that $\bar{\mathcal{A}}^+_b \in \mathrm{LPD}(\Sigma^{2N}_{2s}, \bar{\boldsymbol{\Phi}}, \nu)$ involves $\mathcal{A}^+_b \in \mathrm{LPD}(\bar{\Sigma}^N_s, \boldsymbol{\Phi}, 4\nu)$, with $\bar{\mathcal{A}}^+_b(\boldsymbol{u}) := \mathcal{Q}(\bar{\boldsymbol{\Phi}}\boldsymbol{u} + \boldsymbol{\xi})$, $\bar{\boldsymbol{\Phi}} := (\boldsymbol{\Phi}_{\mathrm{R}}, \boldsymbol{\Phi}_{\mathrm{I}}) \in \mathbb{R}^{m \times 2N}$ and $\Sigma^{2N}_{2s} := \bar{\Sigma}^{2N}_{2s} \cap \mathbb{R}^{2N}$. Indeed, for all $\boldsymbol{u} \in \bar{\Sigma}^N_s$, defining $\boldsymbol{u}^{\mathrm{R}} := (\boldsymbol{u}_{\mathrm{R}}^\top, -\boldsymbol{u}_{\mathrm{I}}^\top)^\top \in \Sigma^{2N}_{2s}$ and $\boldsymbol{u}^{\mathrm{I}} := (\boldsymbol{u}_{\mathrm{I}}^\top, \boldsymbol{u}_{\mathrm{R}}^\top)^\top \in \Sigma^{2N}_{2s}$, we have $\boldsymbol{\Phi}\boldsymbol{u} = \bar{\boldsymbol{\Phi}}\boldsymbol{u}^{\mathrm{R}} + \mathrm{i}\,\bar{\boldsymbol{\Phi}}\boldsymbol{u}^{\mathrm{I}}$ and $\mathcal{A}^+_b(\boldsymbol{u}) = \bar{\mathcal{A}}^+_b(\boldsymbol{u}^{\mathrm{R}}) + \mathrm{i}\,\bar{\mathcal{A}}^+_b(\boldsymbol{u}^{\mathrm{I}})$. Therefore, if $\bar{\mathcal{A}}^+_b \in \mathrm{LPD}(\Sigma^{2N}_{2s}, \bar{\boldsymbol{\Phi}}, \nu)$, $|\langle \mathcal{A}^+_b(\boldsymbol{w}), \boldsymbol{\Phi}\boldsymbol{v}\rangle - \langle \boldsymbol{\Phi}\boldsymbol{w}, \boldsymbol{\Phi}\boldsymbol{v}\rangle| \leq \sum_{r,t \in \{\text{“R”},\text{“I”}\}} |\langle \bar{\mathcal{A}}^+_b(\boldsymbol{w}^r), \bar{\boldsymbol{\Phi}}\boldsymbol{v}^t\rangle - \langle \bar{\boldsymbol{\Phi}}\boldsymbol{w}^r, \bar{\boldsymbol{\Phi}}\boldsymbol{v}^t\rangle| \leq 4\nu m$.

Interestingly, provided $\frac{1}{\sqrt{m}}\|\bar{\boldsymbol{\Phi}}(\boldsymbol{u}-\boldsymbol{v})\| \leq L\eta$ as soon as $\|\boldsymbol{u}-\boldsymbol{v}\| \leq \eta$ for any $\eta > 0$, $L = O(1)$ and $\boldsymbol{u}, \boldsymbol{v} \in \Sigma^{2N}_{2s}$ (*i.e.*, if $\frac{1}{\sqrt{m}}\bar{\boldsymbol{\Phi}}$ is (η, L)-Lipschitz over Σ^{2N}_{2s}), [XJ19, Prop. 6.5] proves that w.h.p. $\bar{\mathcal{A}}^+_b \in \mathrm{LPD}\big(\Sigma^{2N}_{2s}, \bar{\boldsymbol{\Phi}}, \nu = \delta(1+\epsilon)\big)$ provided $m \geq C\delta^{-2} s \ln(\frac{N}{s}) \ln(1 + \frac{c}{\delta^3})$ for some constants $C, c > 0$. However, for all $\boldsymbol{u} := (\boldsymbol{u}_1^\top, \boldsymbol{u}_2^\top)^\top, \boldsymbol{v} := (\boldsymbol{v}_1^\top, \boldsymbol{v}_2^\top)^\top \in \Sigma^{2N}_{2s}$, if $\frac{1}{\sqrt{m}}\boldsymbol{\Phi} \in \mathrm{RIP}(\bar{\Sigma}^N_{2s}, \delta)$, then $\frac{1}{2}\|\bar{\boldsymbol{\Phi}}(\boldsymbol{u}-\boldsymbol{v})\|^2 \leq \|\boldsymbol{\Phi}_{\mathrm{R}}(\boldsymbol{u}_1 - \boldsymbol{v}_1)\|^2 + \|\boldsymbol{\Phi}_{\mathrm{I}}(\boldsymbol{u}_2 - \boldsymbol{v}_2)\|^2 = \|\boldsymbol{\Phi}(\boldsymbol{u}_1 - \boldsymbol{v}_1)\|^2 + \|\boldsymbol{\Phi}(\boldsymbol{u}_2 - \boldsymbol{v}_2)\|^2 \leq m\,(1+\delta)\|\boldsymbol{u}-\boldsymbol{v}\|^2$ since $\boldsymbol{u}_i, \boldsymbol{v}_i \in \Sigma^N_{2s} \subset \bar{\Sigma}^N_{2s}$, which shows that $\frac{1}{\sqrt{m}}\bar{\boldsymbol{\Phi}}$ is $(\eta, 4)$-Lipschitz over Σ^{2N}_{2s} for any $\eta > 0$. This concludes the proof. □

Part III

Phase-Only Acquisition and 1-bit Quantization with Multiplicative Dithering

Chapter 6

Phase-Only Acquisition as an Extension of 1-bit Quantization

This chapter analyses the performances of a simple reconstruction method, namely the Projected Back-Projection (PBP), for estimating the direction of a sparse signal from its phase-only (or amplitude-less) complex Gaussian random measurements, *i.e.*, an extension of one-bit compressive sensing to the complex field. To study the performances of this algorithm, we show that complex Gaussian random matrices respect, with high probability, a variant of the Restricted Isometry Property (RIP) relating to the ℓ_1-norm of the sparse signal measurements to their ℓ_2-norm. This property allows us to upper-bound the reconstruction error of PBP in the presence of phase noise. Monte Carlo simulations are performed to highlight the performance of our approach in this phase-only acquisition model when compared to error achieved by PBP in classical compressive sensing.

6.1 Introduction

One aspect of compressive sensing (CS) is to reduce the number of measurements needed to achieve (high) quality reconstruction of low-complexity signals (*e.g.*, sparse) [Don06a; CRT06b]. Recent research has also focused on reducing the accuracy of each measurement, *e.g.*, by lowering their resolution (or bit-depth) in specific quantization contexts [BB08; Gün+10; Jac+13; JHF11] (see Part II of this thesis). This chapter investigates the consequences of removing the information about the amplitude of a complex signal, *i.e.*, using only the measurement phase for the reconstruction. While phase-only (PO) acquisition can serve as a stepping stone to study new quantizations schemes, *e.g.*, when quantizing the measurement phase [Bou13b] and as developed in the next chapter (see Chap. 7 , this sensing is tantamount to a complex form of one-bit quantization, that was extensively studied in one-bit CS [Fou17; Jac+13; BB08]. The next chapter of this thesis will present how this more abstract acquisition procedure can be linked to applied scenarios.

Oppenheim and co-authors [OL81; OHL82] proved in a few seminal contributions that real, band-limited signals can be reconstructed, up to a lost amplitude, from the phase of their Fourier transform. More recently, for phase-only CS (PO-CS) with complex Gaussian random matrices, Boufounos determined that a specific distance between the measurement phases of two sparse signals encodes their angular distance up to an additive distortion [Bou13a]. While this distortion prevents us from proving perfect estimation of sparse signal direction, the author showed experimentally that this is achievable, thanks to a greedy algorithm enforcing the phase consistency between the signal estimate and the PO measurements.

In this context, our contributions are as follows. While the question of perfect recovery of signal direction remains open, we here focus on a simple, non-iterative algorithm, the Projected Back-Projection (PBP, see Sec. 6.3), and show that this method accurately estimates the direction of sparse signals in PO-CS (Sec. 6.4). This is possible if the sensing matrix respects a variant of the RIP, the (ℓ_1, ℓ_2)-RIP in the complex field, which was previously introduced for (real) one-bit CS. Using tools from measure concentration [Led91], we then prove that complex Gaussian random ma-

trices satisfy, with high probability (w.h.p.), the (ℓ_1, ℓ_2)-RIP if the number of measurements is large compared to the signal sparsity level (Sec. 6.5). Note that the ℓ_1-norm of this RIP prevents a simple proof of this result by recasting the complex field to the real field. Finally, extensive Monte Carlo simulations confirm that the PBP estimation error for PO-CS compares favourably to the one of an unaltered, linear CS scheme (Sec. 6.6).

6.2 Notations and conventions

The Hadamard product is $\odot$; and the angle operator (applied component-wise onto vectors) reads $\angle(ce^{\mathsf{i}\,\phi}) = \phi$ for $c > 0$ and $\phi \in [-\pi, \pi]$. We denote by $\mathcal{N}^{m\times N}(\mu, \sigma^2)$ and $\mathbb{C}\mathcal{N}^{m\times N}(\mu, 2\sigma^2)$ (dropping the symbol N if $N = 1$) the $m \times N$ random matrices with entries independently and identically distributed (i.i.d.) as the normal distribution $\mathcal{N}(\mu, \sigma^2)$ and the complex normal distribution $\mathbb{C}\mathcal{N}(\mu, 2\sigma^2) \sim \mathcal{N}(\mu^{\Re}, \sigma^2) + \mathsf{i}\,\mathcal{N}(\mu^{\Im}, \sigma^2)$, respectively, for some mean μ and variance σ^2. Given $g, g' \sim \mathcal{N}(0, \sigma^2)$, the random variable (*r.v.*) $z := |g + \mathsf{i}\,g'|$ is distributed as the Rayleigh distribution $\mathcal{R}(\sigma)$ with parameter σ [Pap02].

6.3 Phase-Only sensing model

Let us consider a complex s-sparse vector $\boldsymbol{x} \in \bar{\Sigma}_s^N$. Given a complex matrix $\boldsymbol{\Phi} \in \mathbb{C}^{m\times N}$, this work is concerned with the following noisy non-linear sensing model [Bou13a], which generalizes one-bit CS [Fou17; XJ19] to the complex field:

$$\boldsymbol{z} = \operatorname{sign}_{\mathbb{C}}(\boldsymbol{\Phi}\boldsymbol{x}) \odot e^{\mathsf{i}\,\boldsymbol{\xi}}, \tag{6.1}$$

where $\operatorname{sign}_{\mathbb{C}}(\cdot)$ is the *complex signum* operator, applied component-wise onto vectors, *i.e.*, $\operatorname{sign}_{\mathbb{C}}(\lambda) = \lambda/|\lambda|$ for $\lambda \in \mathbb{C} \setminus \{0\}$, and $\boldsymbol{\xi}$ stands for a possible corruption of the measurement phase (with $\xi_i \in [0, 2\pi)$, $i \in [m]$). The matrix $\boldsymbol{\Phi}$ can be, *e.g.*, a complex Gaussian random matrix (see Sec. 6.5).

The sensing model (6.1) thus discards the amplitudes of the measurements $\boldsymbol{\Phi}\boldsymbol{x}$; estimating $\boldsymbol{x}$ from $\boldsymbol{z}$ is possible only up to a global unknown normalization of $\boldsymbol{x}$, *i.e.*, only the direction $\boldsymbol{x}/\|\boldsymbol{x}\|_2$ can be estimated.

We aim to show that the projected back projection (PBP) algorithm

[XJ19; Fou17] accurately estimates the direction of complex sparse signals provided the complex sensing matrix respects a variant of the RIP property (see Sec. 6.4). Given $s \in [N]$, the sensing matrix $\boldsymbol{\Phi}$, and the measurement vector $\boldsymbol{z}$, this algorithm is simply defined as

$$\hat{\boldsymbol{x}} = \mathsf{H}_s\big(\boldsymbol{\Phi}^H \boldsymbol{z}\big), \tag{PBP}$$

where $\mathsf{H}_s(\boldsymbol{u})$ is the hard thresholding operator setting all of the components of the vector $\boldsymbol{u}$ to zero but the s strongest in amplitude (which are unchanged). For CS, PBP is often used as the first iteration of more complex iterative methods such as iterative hard thresholding (IHT) [BD09; XJ19]. Despite its simplicity, analyzing PBP can thus lead to better iterative reconstruction algorithms for PO-CS. The simulations in Chapter 7 showcase the results one can obtain when using PO measurements in conjunction with an adapted QIHT algorithm.

6.4 Bound on the PBP reconstruction error

In CS theory, the error of most signal reconstruction algorithms is controlled by the restricted isometry property — or (ℓ_2, ℓ_2)-RIP — of the sensing matrix [FR13].) We repeat here the definition provided in Sec. 3.2.2, the RIP amounts to asking that for some $\delta > 0$,

$$(1-\delta)\|\boldsymbol{x}\|_2^2 \leq \|\boldsymbol{\Phi}\boldsymbol{x}\|_2^2 \leq (1+\delta)\|\boldsymbol{x}\|_2^2,$$

holds true for all sparse vectors $\boldsymbol{x}$. For instance, if the (real or complex) matrix $\boldsymbol{\Phi}$ respects the (ℓ_2, ℓ_2)-RIP over all $2s$-sparse vectors and one observes a s-sparse vector from the model $\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x}$, as seen in Theorem 3.4, the error of the estimate $\hat{\boldsymbol{x}} = \mathsf{H}_s(\boldsymbol{\Phi}^H \boldsymbol{y})$ is bounded as $\|\boldsymbol{x} - \hat{\boldsymbol{x}}\| = O(\delta)$ [FR13; XJ19] (see Thm. 3.4).

As will be clear below, the capacity of PBP to estimate a sparse vector $\boldsymbol{x}$ from its complex, phase-only observations $\boldsymbol{z}$ in (6.1) depends on the following RIP variant.

Definition 6.1. Given $\delta > 0$, the matrix $\boldsymbol{\Phi} \in \mathbb{C}^{m \times n}$ satisfies the (ℓ_1, ℓ_2)-RIP(s, δ) if, for all $\boldsymbol{x} \in \bar{\Sigma}_s^n$,

$$(1 - \delta)\|\boldsymbol{x}\|_2 \leq \|\boldsymbol{\Phi}\boldsymbol{x}\|_1 \leq (1 + \delta)\|\boldsymbol{x}\|_2.$$

This property was introduced for real one-bit CS [Fou17; PV14]; with it, specific algorithms (including PBP) yield a good estimate of a real sparse signal from the sign of its random measurements. Moreover, provided that m is large compared to s, different types of real random matrix constructions, such as Gaussian random matrices [PV14, Lemma 2.1][JHF11] or randomly sub-sampled Gaussian circulant matrices [DJR19], have been shown to respect the (ℓ_1, ℓ_2)-RIP(s, δ) w.h.p..

To bound the reconstruction error of PBP, we first need the following lemma that is adapted from [Fou17, Lemma 3].

Lemma 6.1. *If $\boldsymbol{\Phi}$ satisfies the (ℓ_1, ℓ_2)-RIP(δ, s) for $0 < \delta < 1$ and $s \in [N]$, then for any vector $\boldsymbol{x} \in \mathbb{C}^N$ with unit ℓ_2-norm such that* $\operatorname{supp} \boldsymbol{x} \subset \mathcal{S} \subset [N]$ *with $|\mathcal{S}| = s$,*

$$\left\| \mathsf{H}_{\mathcal{S}}(\boldsymbol{\Phi}^H \operatorname{sign}_{\mathbb{C}}(\boldsymbol{\Phi}\boldsymbol{x})) - \boldsymbol{x} \right\|_2 \leq \sqrt{5\delta}.$$

We can now determine the main result of this section, which derives from an adaptation of [Fou17, Thm 8] to the complex field.

Theorem 6.2. *If $\boldsymbol{\Phi}$ satisfies (ℓ_1, ℓ_2)-RIP$(2s,\delta)$, then the* PBP *estimate $\hat{\boldsymbol{x}}$ of any signal $\boldsymbol{x} \in \bar{\Sigma}_s^N$ with $\|\boldsymbol{x}\|_2 = 1$ observed via (6.1) with $\|\boldsymbol{\xi}\|_\infty \leq \tau$ respects*

$$\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2 \leq 2\sqrt{5\delta} + 4\tau. \tag{6.2}$$

Proof. Let $\mathcal{S}_0$ and $\mathcal{T}$ be the s-sparse supports of $\boldsymbol{x}$ and $\hat{\boldsymbol{x}}$, respectively. Writing $\mathcal{S} := \mathcal{S}_0 \cup \mathcal{T}$ (with $|\mathcal{S}| \leq 2s$) and $\boldsymbol{a} = \boldsymbol{\Phi}^H \boldsymbol{z}$, we first note that $\|\boldsymbol{x} -$

$\hat{\boldsymbol{x}}\|_2 \leq \|\boldsymbol{x}-\mathsf{H}_{\mathcal{S}}(\boldsymbol{a})\|_2+\|\hat{\boldsymbol{x}}-\mathsf{H}_{\mathcal{S}}(\boldsymbol{a})\|_2$, so that $\|\boldsymbol{x}-\hat{\boldsymbol{x}}\|_2 \leq 2\|\boldsymbol{x}-\mathsf{H}_{\mathcal{S}}(\boldsymbol{a})\|_2$ since $\hat{\boldsymbol{x}}$ is the best s-term approximation of both $\boldsymbol{a}$ and $\mathsf{H}_{\mathcal{S}}(\boldsymbol{a})$. The triangular inequality and Lemma 6.1 then provide

$$\begin{aligned}\|\boldsymbol{x} - \mathsf{H}_{\mathcal{S}}(\boldsymbol{a})\|_2 &= \|\boldsymbol{x} - \mathsf{H}_{\mathcal{S}}\big(\boldsymbol{\Phi}^H[\mathrm{sign}_{\mathbb{C}}(\boldsymbol{\Phi}\boldsymbol{x}) \odot \exp(\mathsf{i}\,\boldsymbol{\xi})]\big)\|_2 \\ &\leq \sqrt{5\delta} + \|\mathsf{H}_{\mathcal{S}}\big(\boldsymbol{\Phi}^H[\mathrm{sign}_{\mathbb{C}}(\boldsymbol{\Phi}\boldsymbol{x}) \odot (1 - e^{\mathsf{i}\,\boldsymbol{\xi}})]\big)\|_2.\end{aligned}$$

Since $\boldsymbol{\Phi}$ respects the (ℓ_1, ℓ_2)-RIP$(2s,\delta)$, we get

$$\begin{aligned}&\|\mathsf{H}_{\mathcal{S}}\big(\boldsymbol{\Phi}^H[\mathrm{sign}_{\mathbb{C}}(\boldsymbol{\Phi}\boldsymbol{x}) \odot (1 - e^{\mathsf{i}\,\boldsymbol{\xi}})]\big)\|_2 \\ &\quad= \sup_{\boldsymbol{u}\in\mathbb{B}^N} \langle \boldsymbol{\Phi}(\mathsf{H}_{\mathcal{S}}(\boldsymbol{u})), \mathrm{sign}_{\mathbb{C}}(\boldsymbol{\Phi}\boldsymbol{x}) \odot (1 - e^{\mathsf{i}\,\boldsymbol{\xi}})\rangle \\ &\quad\leq \|1 - e^{\mathsf{i}\,\boldsymbol{\xi}}\|_\infty \sup_{\boldsymbol{u}\in\mathbb{B}^n} \|\boldsymbol{\Phi}(\mathsf{H}_{\mathcal{S}}(\boldsymbol{u}))\|_1 \\ &\quad\leq 2\|1 - e^{\mathsf{i}\,\boldsymbol{\xi}}\|_\infty \leq 2\|\boldsymbol{\xi}\|_\infty \leq 2\tau.\end{aligned}$$

Gathering all bounds provides the result. □

Interestingly, (6.2) shows that one can still accurately estimate the direction of a complex sparse signal in PO-CS if $\boldsymbol{\Phi}$ is (ℓ_1, ℓ_2)-RIP$(2s, \delta)$ with a small constant δ.

Moreover, as clarified in Sec. 6.5, (6.2) allows us to understand how, for complex Gaussian sensing matrices, the error of PBP decays when m increases. Indeed, up to some missing log factors, we prove in Thm. 6.7 that complex Gaussian random matrices satisfy the (ℓ_1, ℓ_2)-RIP$(2s, \delta)$ w.h.p. provided $m \geq C\delta^{-2}s$ for some $C > 0$. By saturating this condition, we see that, for noiseless PO-CS, PBP achieves the error

$$\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2 = O\big(\sqrt[4]{s/m}\big) \tag{6.3}$$

when m increases, *i.e.*, which tends to zero for large m.

This evolution of the PBP error meets the one encountered for real one-bit CS [Fou17] and non-linear CS [PV16]. However, this behavior is a bit pessimistic compared to the experimental decay in $O(\sqrt{s/m})$ reached by simulations (see Sec. 6.6). The exponent over δ in (6.2) could thus be improved from $\sqrt{\delta}$ to δ. This would then match the performances of PBP in linear CS (see the beginning of this section and Theorem 3.4) and dithered

quantized CS [XJ19; JC17] where it reaches an error bounded by $O(\delta)$ for (ℓ_2, ℓ_2)-RIP$(2s, \delta)$ sensing matrices, *i.e.*, a decay in $O(\sqrt{s/m})$ for Gaussian random sensing matrices.

6.5 The (ℓ_1, ℓ_2)-RIP of Complex Gaussian Matrices

While one easily extends the (ℓ_2, ℓ_2)-RIP of certain random matrix constructions from the real to the complex fields — *e.g.*, by recasting the signal space $\mathbb{C}^N$ and measurement domain $\mathbb{C}^m$ to $\mathbb{R}^{2N}$ and $\mathbb{R}^{2m}$, respectively [FR13] — such an extension for (ℓ_1, ℓ_2)-RIP matrices is not known.

Fortunately, using the tools of measure concentration [Led91], we prove below that complex Gaussian random matrices $\boldsymbol{\Phi}$ respects the (ℓ_1, ℓ_2)-RIP w.h.p. provided m is large compared to the signal sparsity. To show this, we first establish that, given $\boldsymbol{x} \in \mathbb{C}^N$, $\mathbb{E}\|\boldsymbol{\Phi}\boldsymbol{x}\|_1$ is proportional to $\|\boldsymbol{x}\|_2$ since each random variable $|(\boldsymbol{\Phi}\boldsymbol{x})_i|$ is Rayleigh distributed.

Lemma 6.3. *Given $\boldsymbol{x} \in \mathbb{C}^N$ and a random matrix $\boldsymbol{\Phi} \sim \mathbb{C}\mathcal{N}^{m\times n}(0, \sigma^2)$ with $\sigma := \frac{1}{m}\frac{\sqrt{2}}{\sqrt{\pi}}$, we have*

$$\mathbb{E}\big[\|\boldsymbol{\Phi}\boldsymbol{x}\|_1\big] = \|\boldsymbol{x}\|_2.$$

Proof. By decomposing both the entries of $\boldsymbol{\Phi}$ and the components of $\boldsymbol{x}$ into their real and imaginary parts, we get

$$\|\boldsymbol{\Phi}\boldsymbol{x}\|_1 = \sum_{i=1}^{m} \big|\sum_{j=1}^{N} \Phi_{ij}x_j\big| = \sum_{i=1}^{m} \big|\sum_{j=1}^{n} g_{ij}^{\Re} + \mathsf{i}\, g_{ij}^{\Im}\big|,$$

with $g_{ij}^{\Re} := \Phi_{ij}^{\Re}x_j^{\Re} - \Phi_{ij}^{\Im}x_j^{\Im}$ and $g_{ij}^{\Im} := \Phi_{ij}^{\Re}x_j^{\Im} + \Phi_{ij}^{\Im}x_j^{\Re}$.

We note that, for all indices $i, i' \in [m]$ and $j, j' \in [N]$, $\Phi_{ij}^{\Re}$ and $\Phi_{ij}^{\Im}$ are Gaussian random variables with $\mathbb{E}[\Phi_{ij}^{\Re}] = \mathbb{E}[\Phi_{ij}^{\Im}] = \mathbb{E}[\Phi_{ij}^{\Re}\Phi_{i'j'}^{\Im}] = 0$. Therefore, $g_{ij}^{\Re}, g_{ij}^{\Im} \sim_{\text{i.i.d.}} \mathcal{N}(0, \sigma^2|x_j|^2)$ and a simple computation provides $\mathbb{E}g_{ij}^{\Re}g_{i'j'}^{\Im} = 0$. The *r.v.*s $\Gamma_i^{\Re} := \sum_{j=1}^{n} g_{ij}^{\Re}$ and $\Gamma_i^{\Im} := \sum_{j=1}^{n} g_{ij}^{\Im}$ are thus

independent and distributed as $\mathcal{N}(0, \sigma^2\|\boldsymbol{x}\|_2^2)$ for all $i \in [m]$. Consequently,

$$\mathbb{E}\big[\|\boldsymbol{\Phi}\boldsymbol{x}\|_1\big] = \sum_{i=1}^{m} \mathbb{E}\big[|\Gamma_i^{\Re} + \mathrm{i}\,\Gamma_i^{\Im}|\big] = m\mathbb{E}\big[\Gamma_0\big],$$

where Γ_0 follows a Rayleigh distribution $\mathcal{R}(\sigma\|\boldsymbol{x}\|_2)$. Since $\mathbb{E}[\Gamma_0] = \sigma\sqrt{\frac{\pi}{2}}\|\boldsymbol{x}\|_2$ [Pap02] and $\sigma = \frac{1}{m}\sqrt{\frac{2}{\pi}}$, we find $\mathbb{E}\big[\|\boldsymbol{\Phi}\boldsymbol{x}\|_1\big] = \sigma\|\boldsymbol{x}\|_2\sqrt{\frac{\pi}{2}}m = \|\boldsymbol{x}\|_2$. □

The following proof uses ρ-covering of the set of complex s-sparse vectors $\tilde{\Sigma}_s^N$.

Definition 6.2. A covering $\mathcal{J}_\rho$ is defined as the union of points $\boldsymbol{u}_j \in \mathcal{J}_\rho \subset \tilde{\Sigma}_s^N$ whose associated ℓ_2-ball of radius ρ can cover entirely the set $\tilde{\Sigma}_s^N$ *i.e.*,

$$\tilde{\Sigma}_s^N \subset \cup_j^{|\mathcal{J}_\rho|}\{\boldsymbol{x} \in \tilde{\Sigma}_s^N, \|\boldsymbol{x} - \boldsymbol{u}_j\|_2 \leq \rho\}.$$

This covering has a bounded size.

Lemma 6.4 (ρ-covering of $\tilde{\Sigma}_s^N$)**.** *A ρ-covering $\mathcal{J}_\rho$ that covers the set of s-sparse vectors $\tilde{\Sigma}_s^N$ has a size that can be upper-bounded by*

$$|\mathcal{J}_\rho| \leq \binom{N}{s}(1 + \frac{2}{\rho})^{2s} \leq (\frac{eN}{s})^s(1 + \frac{2}{\rho})^{2s}.$$

Proof. We note that $\tilde{\Sigma}_s^N = \bigcup_{\mathcal{S}\subset[N]:|\mathcal{S}|=s} \tilde{\Sigma}^N(\mathcal{S})$, with $\tilde{\Sigma}^N(\mathcal{S}) := \{\boldsymbol{u} \in \bar{\mathbb{B}}^N : \mathrm{supp}(\boldsymbol{u}) = \mathcal{S}\}$. Moreover, $\tilde{\Sigma}^N(\mathcal{S})$ is isomorphic to $\bar{\mathbb{B}}^s$, and thus to $\mathbb{B}^{2s}$. Since this last set, and thus $\tilde{\Sigma}^N(\mathcal{S})$, can be covered with no more than $(1 + \frac{2}{\rho})^{2s}$ vectors [Bar+08], a covering $\mathcal{J}_\rho$ of $\tilde{\Sigma}_s^N$ can be reached by gathering all coverings — $\binom{N}{s}$ in total — so that

$$|\mathcal{J}_\rho| \leq \binom{N}{s}(1 + \frac{2}{\rho})^{2s} \leq (\frac{eN}{s})^s(1 + \frac{2}{\rho})^{2s}.$$

□

Remark 6.1. Interestingly, by design, this covering is such that all $\boldsymbol{x} \in \tilde{\Sigma}_s^N$ can be written as $\boldsymbol{x} = \boldsymbol{u} + \boldsymbol{r}$ with $\boldsymbol{u} \in \mathcal{J}_\rho \subset \tilde{\Sigma}_s^N$, $\boldsymbol{r} \in \rho\bar{\mathbb{B}}^N \cap \tilde{\Sigma}_s^N = \rho\tilde{\Sigma}_s^N$, with $\operatorname{supp}\boldsymbol{x} = \operatorname{supp}\boldsymbol{u} = \operatorname{supp}\boldsymbol{r}$.

We also need this classical result from Ledoux and Talagrand [Led91, Eq. 1.6], see also [JHF11, Lemma 5].

Lemma 6.5. *If the function F is Lipschitz with $\lambda = \|F\|_{Lip}$, then, for $r > 0$ and $\boldsymbol{\gamma} \sim \mathcal{N}^m(0,1)$,*

$$\mathbb{P}\big(\big|F(\boldsymbol{\gamma}) - \mathbb{E}(F(\boldsymbol{\gamma}))\big| > r\big) \leq 2\exp(-\frac{1}{2}r^2\lambda^{-2}).$$

In our developments, F will be of the following kind.

Lemma 6.6. *The functions $G : \boldsymbol{u} \in \mathbb{C}^m \mapsto \|\boldsymbol{u}\|_1 \in \mathbb{R}_+$ and of $G' : (\boldsymbol{u}^\Re, \boldsymbol{u}^\Im) \in \mathbb{R}^{m\times 2} \mapsto \|(\boldsymbol{u}^\Re, \boldsymbol{u}^\Im)\|_{2,1} \in \mathbb{R}_+$ have a Lipschitz constant equal to $\sqrt{m}$.*

Proof. For all $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{C}^m$, $|\|\boldsymbol{u}\|_1 - \|\boldsymbol{v}\|_1| \leq \|\boldsymbol{u} - \boldsymbol{v}\|_1 \leq \sqrt{m}\|\boldsymbol{u} - \boldsymbol{v}\|_2$, which gives the Lipschitz constant of G. The one of G' follows from $\|\boldsymbol{u}\|_1 = \|(\boldsymbol{u}^\Re, \boldsymbol{u}^\Im)\|_{2,1}$. □

We are now ready to prove the main result of this section.

Theorem 6.7. *Let $\delta \in (0,1)$, $\sigma = \frac{1}{m}\frac{\sqrt{2}}{\sqrt{\pi}}$, and $\boldsymbol{\Phi} \sim \mathbb{C}\mathcal{N}^{m\times N}(0, \sigma^2)$ be a complex Gaussian random matrix. If $m \geq \frac{36}{\pi}\delta^{-2}\big[s\log\big(\frac{eN}{s}(1+\frac{6}{\delta})^2\big) + \log(\frac{2}{\eta})\big]$, then, with probability exceeding $1-\eta$, the matrix $\boldsymbol{\Phi}$ satisfies the (ℓ_1, ℓ_2)-RIP(s, δ).*

Proof. The proof strategy follows the one developed in [Bar+08] for proving that real Gaussian random matrices satisfy the (ℓ_2, ℓ_2)-RIP w.h.p.. By homogeneity of the (ℓ_1, ℓ_2)-RIP, it is enough to prove that complex Gaussian random matrices satisfy it w.h.p. for all vectors of $\tilde{\Sigma}_s^N := \bar{\Sigma}_s^N \cap \bar{\mathbb{B}}^N$.

We first show that for a fixed vector $\boldsymbol{x} \in \mathbb{C}^N$, $\|\boldsymbol{\Phi x}\|_1$ concentrates around $\|\boldsymbol{x}\|_2$. Using the *r.v.*s $\Gamma_i^{\Re}, \Gamma_i^{\Im}$ defined in the proof of Lemma 6.3, we can write

$$\begin{aligned} p &:= \mathbb{P}\big(\big| \|\boldsymbol{\Phi x}\|_1 - \|\boldsymbol{x}\|_2 \big| > t\|\boldsymbol{x}\|_2 \big) \\ &= \mathbb{P}\big(\big| \sum_{i=1}^m \big((\Gamma_i^{\Re})^2 + (\Gamma_i^{\Im})^2 \big)^{1/2} - \|\boldsymbol{x}\|_2 \big| > t\|\boldsymbol{x}\|_2 \big) \\ &= \mathbb{P}\big(\big| \sum_{i=1}^m \big((\gamma_i^{\Re})^2 + (\gamma_i^{\Im})^2 \big)^{1/2} - m\sqrt{\frac{\pi}{2}} \big| > tm\sqrt{\frac{\pi}{2}} \big), \end{aligned}$$

where we defined the independent Gaussian random vectors $\boldsymbol{\gamma}^{\Re}, \boldsymbol{\gamma}^{\Im} \sim_{\text{i.i.d.}} \mathcal{N}^m(0,1)$. Since $\sum_{i=1}^m \big((\gamma_i^{\Re})^2 + (\gamma_i^{\Im})^2 \big)^{1/2} = \|(\boldsymbol{\gamma}^{\Re}, \boldsymbol{\gamma}^{\Im})\|_{2,1}$, Lemma 6.5 provides

$$\begin{aligned} p &= \mathbb{P}\big(\big| \, \|(\boldsymbol{\gamma}^{\Re}, \boldsymbol{\gamma}^{\Im})\|_{2,1} - m\sqrt{\frac{\pi}{2}} \big| > tm\sqrt{\frac{\pi}{2}} \big) \\ &\leq 2\exp\big(-\frac{\pi}{4} t^2 m \big) \end{aligned} \tag{6.4}$$

by considering $\boldsymbol{\gamma} = (\boldsymbol{\gamma}^{\Re}, \boldsymbol{\gamma}^{\Im})$ as a $2m$ Gaussian random vector, with the function $F(\boldsymbol{\gamma}) := \|(\boldsymbol{\gamma}^{\Re}, \boldsymbol{\gamma}^{\Im})\|_{2,1}$ whose Lipschitz constant is characterized in Lemma 6.6. Therefore, given $\boldsymbol{x}$ and $t > 0$, we have

$$\big| \|\boldsymbol{\Phi x}\|_1 - \|\boldsymbol{x}\|_2 \big| \leq t\|\boldsymbol{x}\|_2,$$

with probability exceeding $1 - p \geq 1 - 2\exp\big(-\frac{\pi}{4} t^2 m \big)$.

We now extend this result to all vectors of $\tilde{\Sigma}_s^N$ by first determining when this concentration holds for all the vectors of a ρ-covering of this domain defined in Lemma 6.4 — that is a set such that all elements of $\tilde{\Sigma}_s^N$ are no more than $\rho > 0$ far apart from an element of this covering — and by finally extending this property to $\tilde{\Sigma}_s^N$ by continuity.

Using (6.4), by union bound over all the vectors of $\mathcal{J}_\rho$, the event

$$\mathcal{E}_{\rho,t} : \quad \big| \|\boldsymbol{\Phi u}\|_1 - \|\boldsymbol{u}\|_2 \big| \;\leq\; t, \quad \forall \boldsymbol{u} \in \mathcal{J}_\rho, \tag{6.5}$$

holds with failure probability $p_{\rho,t} := \mathbb{P}(\mathcal{E}_{\rho,t}^c)$ at most

$$p_{\rho,t} \leq 2\big(\frac{eN}{s}\big)^s \big(1+\frac{2}{\rho}\big)^{2s} \exp\big(-\frac{\pi}{4}t^2 m\big).$$

Let us assume $\mathcal{E}_{\rho,t}$ holds and pick an arbitrary $\boldsymbol{x} \in \tilde{\Sigma}_s^N$. As explained above in Remark 6.1, we can write $\boldsymbol{x} = \boldsymbol{u} + \boldsymbol{r}$ with $\boldsymbol{u} \in \mathcal{J}_\rho$, $\boldsymbol{r} \in \rho\tilde{\Sigma}_s^N$, and $\operatorname{supp}\boldsymbol{x} = \operatorname{supp}\boldsymbol{u} = \operatorname{supp}\boldsymbol{r}$.

Using (6.5), and the properties of the covering, we get

$$\begin{aligned}
&|\|\boldsymbol{\Phi}\boldsymbol{x}\|_1 - \|\boldsymbol{x}\|_2| = |\|\boldsymbol{\Phi}(\boldsymbol{u}+\boldsymbol{r})\|_1 - \|(\boldsymbol{u}+\boldsymbol{r})\|_2| \\
&\leq |\|\boldsymbol{\Phi}\boldsymbol{u}\|_1 - \|\boldsymbol{u}\|_2| + |\|\boldsymbol{\Phi}(\boldsymbol{u}+\boldsymbol{r})\|_1 - \|\boldsymbol{\Phi}\boldsymbol{u}\|_1| \\
&+ |\|\boldsymbol{u}+\boldsymbol{r}\|_2 - \|\boldsymbol{u}\|_2| \leq t + \rho + \rho\|\boldsymbol{\Phi}(\rho^{-1}\boldsymbol{r})\|_1,
\end{aligned}$$

where we used multiple times the triangular inequality. However, $\rho^{-1}\boldsymbol{r} \in \tilde{\Sigma}_s^N$ and we can recursively apply the same development to $\|\boldsymbol{\Phi}(\rho^{-1}\boldsymbol{r})\|_1$, so that

$$|\|\boldsymbol{\Phi}\boldsymbol{x}\|_1 - \|\boldsymbol{x}\|_2| \leq (t+\rho)\sum_{k=0}^{+\infty}\rho^k = \frac{t+\rho}{1-\rho}.$$

Setting $t = \rho = \delta/3$ for some $0 < \delta < 1$, we get $\frac{t+\rho}{1-\rho} \leq \delta$. From the analysis of $\mathcal{E}_{\rho,t}$ above, we finally obtain that $|\|\boldsymbol{\Phi}\boldsymbol{x}\|_1 - \|\boldsymbol{x}\|_2| \leq \delta$ holds true for all $\boldsymbol{x} \in \tilde{\Sigma}_s^N$ — *i.e.*, the (ℓ_1,ℓ_2)-RIP is verified — with failure probability at most

$$p_{\frac{\delta}{3},\frac{\delta}{3}} \leq 2\big(\frac{eN}{s}\big)^s \big(1+\frac{6}{\delta}\big)^{2s} \exp\big(-\frac{\pi}{36}\delta^2 m\big).$$

We conclude the proof by observing that $p_{\frac{\delta}{3},\frac{\delta}{3}} \leq \eta$ for $0 < \eta < 1$ if $m \geq \frac{36}{\pi}\delta^{-2}\big[s\log\big(\frac{eN}{s}(1+\frac{6}{\delta})^2\big) + \log(\frac{2}{\eta})\big]$. □

6.6 Simulations

We now assess the tightness of our theoretical analysis through Monte Carlo simulations. We do not aim to demonstrate the superiority of (PBP) over other methods but to study the potentialities of such a simple algorithm in PO-CS.

As a first experiment, we have tested the estimation of complex sparse signals $\boldsymbol{x}$ in $\mathbb{C}^N$ with $N = 256$ for different sparsity levels $s \in [N]$ and

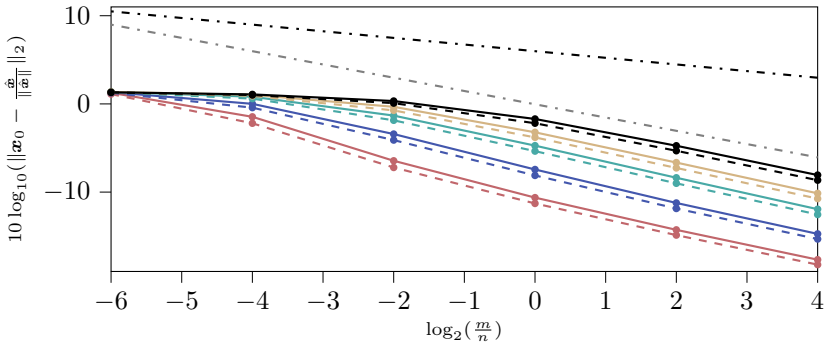


Figure 6.1: (Best viewed in color) Reconstruction error of (PBP) for different measurement models. (dashed lines) compressive sensing; (solid lines) phase-only measurements. The colors represent the sparsity, namely $s = 2$ in **red**, $s = 4$ in **blue**, $s = 10$ in **green**, $s = 20$ in **yellow**, and $s = 50$ in **black**. The dotted lines represent the rates of $m^{-\frac{1}{2}}$ in **gray** and $m^{-\frac{1}{4}}$ in black.

measurement number m. Two acquisition strategies were compared: the phase-only acquisition fixed by the model (6.1), and classical compressive sensing where we directly acquire the measurement vector $\boldsymbol{y} := \boldsymbol{\Phi}\boldsymbol{x}$ without alteration. For each combination of s and m, the performances of both strategies have been tested over 100 000 generations of the sparse signal $\boldsymbol{x}$ and the complex Gaussian random matrix $\boldsymbol{\Phi} \sim \mathbb{C}\mathcal{N}(0, \sigma^2)$, with σ^2 set to $2/(\pi m^2)$ and $1/m$ for the phase-only and the CS scenario, respectively. Each sparse signal $\boldsymbol{x}$ was created by picking a s-sparse support uniformly at random amongst the $\binom{N}{s}$ possible supports, inserting in this support s i.i.d. complex values picked uniformly at random before normalizing. We analyzed the reconstruction error of the signal direction with the metric $\mathcal{E}(\boldsymbol{x}, \hat{\boldsymbol{x}}) := \|\boldsymbol{x} - \|\hat{\boldsymbol{x}}\|_2^{-1}\hat{\boldsymbol{x}}\|_2$, where $\hat{\boldsymbol{x}}$ is the (PBP) estimate. Comparing the two schemes in Fig. 6.1 for different sparsity levels, we observe that the reconstruction error achieved from phase-only measurements exhibits good performances given the absence of the amplitude information. The experimental convergence rate is also matching the one of the CS scheme; it scales as $m^{-\frac{1}{2}}$ when m increases instead of the pessimistic rate in $m^{-\frac{1}{4}}$ predicted by the theory in (6.3). The phase-only scheme seems to only suffer from a constant loss (in dB) when compared to the classic model.

In a second experiment, we have studied the performances of PBP in the presence of phase noise. In this new test, we kept the same parameters as

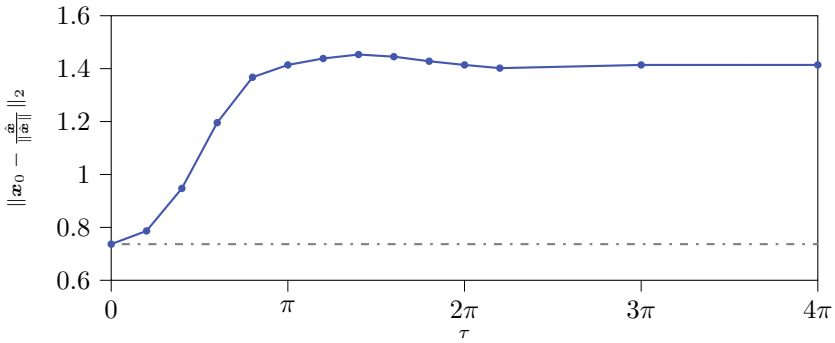


Figure 6.2: Reconstruction error of (PBP) for noiseless (dashed lines) and noisy measurements (solid lines) for different τ with $s = 10$ and $m = 64$.

above, restricting only the sparsity level and the number of measurements to $s = 10$ and $m = 64$, respectively. The phase noise $\boldsymbol{\xi}$ in (6.1) was generated according to a uniform distribution between $-\tau$ and τ, with $\tau \in [0, 4\pi]$. As established (6.2), the reconstruction error $\mathcal{E}(\boldsymbol{x}, \hat{\boldsymbol{x}})$ increases almost linearly when τ increases from 0 to π, before saturating at $\sqrt{2}$ from $\tau > \pi$. In other words, from that noise level, phase-only measurements are too noisy and $\langle \boldsymbol{x}, \hat{\boldsymbol{x}} \rangle \approx 0$. Furthermore, the additive nature of the degradation in (6.2) is clearly visible when comparing the noiseless in dashed gray and noisy reconstruction in solid **blue**.

6.7 Discussion

In this chapter, we have studied how to estimate the direction of complex sparse vectors from noisy phase-only measurements. We proved theoretically that the estimate yielded by the projected back projection of noisy phase-only measurement has bounded and stable reconstruction error provided that the sensing matrix satisfies an extension of the (ℓ_1, ℓ_2)-RIP in the complex field. Moreover, we showed that $m \times N$ complex Gaussian random matrices respect w.h.p. this property with distortion $\delta > 0$ provided that m is large compared to the signal sparsity level s, *i.e.*, $m = O(\delta^{-2} s \log(\frac{N}{\delta s}))$. The proof of this result leverages the tools of measure concentration since the ℓ_1-norm prevents a simple recasting of the complex (ℓ_1, ℓ_2)-RIP to a real domain of larger dimension. We finally analyzed the tightness of our the-

oretical developments through Monte Carlo simulations. They confirmed that, despite the lack of amplitude information, we can reach arbitrary high accuracy on the estimation of sparse signal direction provided m/s is large, with an experimental error rate decaying as $1/\sqrt{m}$ when m increases, thus faster than our theoretical error rate in $1/m^{1/4}$. The discrepancy between this two rates will be studied in future work, as well as the impact of phase quantization and additive noise on the phase-only sensing model.

The results introduced in this chapter show that, even basic algorithm like PBP can have strong theoretical guarantees. In [JF21] these results were extended to the use of *instance optimal* algorithms like BPDN (see Theorem 3.2) and showed that perfect recovery is possible for models relying on complex Gaussian matrices.

Chapter 7

Multiplicative Dithering for 1-bit CS Radar

In this chapter, we tackle the issue of implementing a dithering procedure for the 1-bit quantization of radar signals that is able to generate high quality estimates while remaining a low complexity and cost efficient solution. Specifically, we stray away from the additive dithering that induces, as will be made clear, a complex and high cost implementation. Instead, we propose the use of a multiplicative dithering. This process can leverage already existing radar architecture of Frequency Modulated Continuous Wave (FMCW) radars and can be thus efficiently implemented. The efficiency of this multiplicative dithering is first studied theoretically and its link to another coarse quantization scheme, namely the Phase-Only acquisition, is highlighted. The performances of this novel dithering scheme is then extensively tested using Monte Carlo simulations and is then thoroughly compared to its additive counter-part. A hardware relaxed version of the random phase dithering is also introduced and compared to the other 1-bit schemes. The observations made in simulations are then validated using actual radar measurements at 24GHz. These measurements, combined with the simulations, show that the multiplicative dithering is a good alternative to the additive random dithering in a low number of measurements setting. Specifically, we show that this procedure is a good trade-off between strong theoretical guarantees and reconstruction quality for low complexity

hardware.

7.1 Problem Statement

In applications where the aim of 1-bit quantization is to lower the requirement on the hardware by either lowering the cost or the power consumption, the design of the dithering process is capital. Indeed, if no dithering is applied, the implementation is extremely simplified but at a cost in terms of performances as shown in the previous chapters of this thesis (Chapter 4 and 5) and in [Feu+18b; JC17; PV12; Feu+18a]. In [PV12] for Bernoulli matrices and in [Feu+18a; Feu+18b] for Fourier matrices, authors showed that applying this extremely low resolution quantizer on noiseless data can results in ambiguous scenarios preventing high quality reconstruction. One such scenario arises when two different sparse vectors, once quantized, are sent exactly to the same bits. This removes any possibility to distinguish them after quantization. For example

$$\mathcal{Q}_\epsilon(\boldsymbol{\Phi}\boldsymbol{x}_0) = \mathcal{Q}_\epsilon(\boldsymbol{\Phi}\boldsymbol{x}_1),$$

with $\boldsymbol{x}_0 \neq \boldsymbol{x}_1$. This is one of the reasons that motivated the use of additive dithering in highly quantized applications. This can be modeled by a vector $\boldsymbol{\xi} \in \mathbb{C}^m$ that is added to the signal before the quantization. As seen in Part II and Sec. 3.3.2, the quantizer then becomes :

$$\mathcal{Q}_\epsilon^+(\boldsymbol{y}) := \mathcal{Q}_\epsilon(\boldsymbol{y} + \boldsymbol{\xi}).$$

There exist different ways of generating this dither. Authors in [Kam+12; Dir19; Bar+17] proposed to generate these varying thresholds according to the previous measurements in order to extract the most information possible out of those coarse 1-bit measurements. While this process of generating *tailored* thresholds in a iterative fashion induces a steep performance gain in the reconstruction, this process implies a feedback loop between the reconstruction and the 1-bit acquisition that can be quite expensive, be it in cost or power to implement. Indeed, it requires a generation of the varying thresholds that can be entirely controlled by the processing unit and of high resolution.

Authors in [XJ19], proposed to use a dither that is generated randomly according to a uniform distribution $\boldsymbol{\xi} \sim \mathcal{U}^m(-\frac{\epsilon}{2}, \frac{\epsilon}{2})$. They were able to show that by leveraging the effect of this dither on the quantized data, one can upperbound the ℓ_2-reconstruction of the Projected-Back Projection (PBP) algorithm (Remark 3.3 from [XJ19]). This way of dithering only requires a generation of the dither but no control or influence by the reconstruction algorithm, it is purely random.

In Fig. 7.1, the hardware introduced in Fig. 2.5 is modified for the 1-bit and additively dithered acquisition. It is clear that adding a uniform

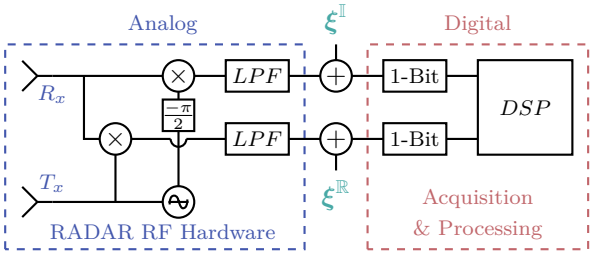


Figure 7.1: Example of FMCW radar architecture with additive dithering and 1-bit quantization

random variable gives interesting recovery guarantees [XJ19] and as seen in Part II. Actually implementing this dither in practice, however, comes with numerous challenges. Indeed, and as stated before, the added dither must follow a uniform distribution that spans the dynamic of the measured signal, *i.e.*, $\|\boldsymbol{y}\|_\infty$. The problems associated to this constraint are twofold. First, one must find a way of generating this uniform random variable in hardware at a cost or power consumption that is still attractive compared to high resolution ADCs. This means either finding a hardware component that is able to generate this dither or using high resolution DACs. One can already notice that trading high resolution ADCs for high resolutions DACs does not really simplify the acquisition process. Second, having this random variable span the dynamic of the signal implies that this dynamic is partially known. This assumption is a rather strong one given the high variability of received powers that can be encountered in radar signals. Indeed, for a constant RCS (radar cross section), the received power varies as $\mathcal{O}(R^{-4})$ in a monostatic radar (see the *radar's equation* (2) in Chapter 2.1). This

means that there is more than a $10dB$ loss of amplitude every time the range of a target doubles. The simulations in Section 6.6 will highlight further the impact of an imperfect estimate or knowledge of this dynamic. The last issue with this architecture is the fact that certain algorithms, such as QIHT introduced in this thesis in (4.10), require the exact knowledge of the dithering that was used in order to impose a consistency condition between the measurement and the estimated signal. This means that the dither must be controlled or measured before or during the acquisition, which imposes further constraint on the hardware.

Given the issues associated with this additive dithering but the need of going beyond the simple deterministic quantization, we set out to find another way of dithering signals coming from FMCW radars. We introduce the multiplicative dither that modifies the phases of the measured signals before the 1-bit quantization according to unitary vector with a random phase.

The contributions of the chapter are the following: *(i)* we show that using a multiplicative dither in the context of 1-bit quantization is a viable trade-off between reconstruction performances and complexity of implementation in hardware; *(ii)* we further show that one can relax the random nature of the phase used in this dither to a structured one without sacrificing too much performances while simplifying even more the hardware implementation of the dithering process; *(iii)* this structured dither is shown to enable the use of consistency based algorithm (like QIHT) without any added complexity that is commonly found in the additive procedure; *(iv)* we show that this multiplicative scheme has reconstruction performances bounded by what is achievable by Phase-Only acquisition; (*v*) similarly to Part II, we show that PO measurement for Fourier based measurements are subject to ambiguities. (*vi*) finally, all of those results are confirmed using Monte Carlo simulations and real radar measurements.

This chapter is structured as follows. Section 7.1 introduces the issue of implementing an efficient dithering procedure. Sec. 7.2 introduces the multiplicative dithering studied in this paper and highlights its advantages compared to the additive dither in terms of practical implementation. Sec. 7.3 links the multiplicative dithering to the PO acquisition and studies its limits in terms of uniform recovery guarantees of Fourier based measurements.

Sec. 7.4 studies a non-uniform ℓ_2-bound on the discrepancy between the reconstruction obtained with this multiplicative dither and the PO-CS acquisition. The results of Monte-Carlo simulations are presented in Sec. 7.5. These performances are then asserted in Section 7.6, where real radar measurements are used. Finally, conclusions and future works are presented in Section 7.7.

7.2 Multiplicative Dithering

In this chapter, instead of dithering the amplitudes of the received signal in the real and imaginary domain, we propose to dither their phases by multiplying the signal with a unitary signal with a random phase. The proposed quantization process can be represented as :

$$\mathcal{Q}^{\odot}(\boldsymbol{y}) = c_{\odot}\mathcal{Q}(\boldsymbol{y} \odot \boldsymbol{\xi}) \odot \boldsymbol{\xi}^*,$$

with $c_{\odot} = \frac{\pi}{4}$. A complete study of the performance of this scheme is presented is Section 7.4.

This way of dithering solves the first issue brought up by the previous section regarding the additive scheme. It is indeed amplitude independent and can be thus applied without the knowledge of the dynamic of the signal $\|\boldsymbol{y}\|_{\infty}$. Furthermore, dithering the phase can be achieved using different and already existing architectures. Using a non *zero-if* demodulation architecture [MSM17] such as in Fig. 7.2. One needs to generate the dither $\xi^*(t)$, the radar transmits the original signal $s(t)$ and then the received RF signal $r^{RF}(t)$is demodulated by $s(t)\xi^*(t)$. The coherent demodulation in (2.4) then becomes

$$r_{n0}(t) = r^{RF}(t)\big(s(t)\xi^*(t)\big)^* = r(t)\xi(t).$$

Sampling the time signal $r_{n0}(t)$, similarly to Part II, finally gives $\boldsymbol{r} \odot \boldsymbol{\xi}$.

Another way of randomly modifying the phase of the received signal is to modify the signal after its coherent demodulation. In Fig. 7.3, the baseband signals are multiplied by the real and imaginary parts of the dither

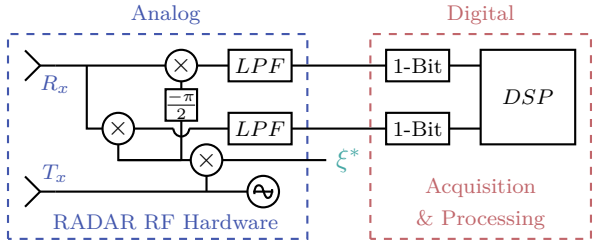


Figure 7.2: Example of FMCW radar architecture with non zero IF demodulation and 1-bit quantization

$\xi^*(t) = \xi^{\mathbb{R}}(t) + \mathsf{i}\,\xi^{\mathbb{I}}(t)$ and combined in the following way,

$$\begin{aligned} r(t) \times \xi(t) =& r^{\mathbb{R}}(t) \times \xi^{\mathbb{R}}(t) - r^{\mathbb{I}}(t) \times \xi^{\mathbb{I}}(t) \\ &+ \mathsf{i}\left(r^{\mathbb{R}}(t) \times \xi^{\mathbb{I}}(t) + r^{\mathbb{I}}(t) \times \xi^{\mathbb{R}}(t)\right). \end{aligned}$$

This process can be easily implemented with *off-the-shelf* components. Indeed, it only requires base-band components operating thus at low frequencies. These are both inexpensive and require low power compared to their RF counterparts. A future publication will study an actual 1-bit radar prototype based on this architecture.

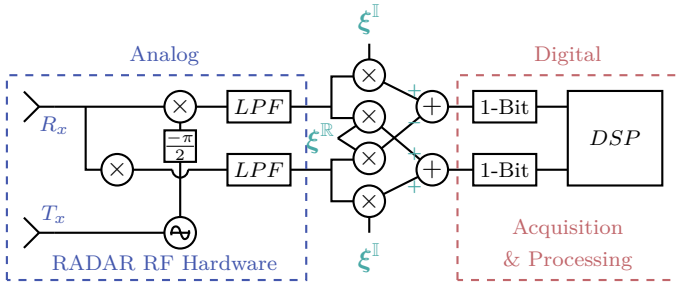


Figure 7.3: Example of FMCW radar architecture with multiplicative dithering and 1-bit quantization

While the application of the dither has been simplified in the architecture found in Fig. 7.2 and Fig. 7.3, there is still some complexity involved in the generation of the dither $\boldsymbol{\xi}$ itself. Indeed, in both dithering methods, be it additive or multiplicative, the dither needs to follow a specific distribution that one must generate before applying it to the radar signals. Although

the multiplicative dither does not require the knowledge of the dynamic of the signal, its random phase still requires the use of high resolution DACs for its generation.

We now propose to relax this constraint on the randomness of the phase of the dither by replacing it by a deterministic function that can be easily implemented in hardware. The unitary dither with a random phase is replaced by

$$\xi^*(t) := \exp\left(\mathrm{i}\,\Delta_f t + \chi\right). \tag{7.1}$$

This tremendously simplifies the hardware implementation. In the context of Fig. 7.3, one only needs to create two signals, namely $\cos(\Delta_f t)$ and $\sin(\Delta_f t)$ using a known frequency Δ_f. Sine and cosine functions are easily generated using based-band components without using high resolution DACs [GTH71].

Aside from the obvious hardware simplification that this structured dither provides, it can also be leveraged in algorithms that enforce a consistency between the measurements and the estimate such as QIHT [JDD13; Feu+18b]. Indeed, only the knowledge of Δ_f and of the sampling times (*e.g.*, t_i in Sec. 4.2) are needed to recreate this dither and use it in the reconstruction. The reconstructed sparse signal will just be phase-shifted by the unknown phase χ, which is inconsequential in most radar applications. This also means that there is no need for a synchronisation or feedback between the dither and the radar and the processing. The multiplicative dithering and its generation act as a completely separate system, simplifying its implementation even further compared to system that requires the control or knowledge of the dither to enforce consistency [Kam+12; Feu+18b]. It is also interesting to note that although this multiplication by a single tone dither might increase the frequency content of acquired signal, its total bandwidth remains unchanged which means the sampling frequency remains constant. This deterministic dithering only works for models where increasing the number of measurements corresponds to acquiring (before the quantization) repetitions of the signals, like in ranging applications with FMCW radars.

While being advantageous in its implementation, using a multiplicative dither, as we will see in the next sections, does not provide as strong theoretical guarantees as the one provided by the additive dithering. While the

performances of the additive dithering are upper-bounded by what can be obtained by classic high resolution measurements as $\mathbb{E}(\mathcal{Q}_\epsilon^+(\boldsymbol{y})) = \boldsymbol{y}$ [XJ19] (see also Lem. 3.6) . Acquiring the measurements using $\mathcal{Q}^\odot(\cdot)$ does not capture all of the information about the signal of interest. This is best exhibited by

Lemma 7.1. *For $\mathcal{Q}^\odot(\cdot)$, any $a \in \mathbb{C}$, we have*

$$\mathbb{E}_\xi\{\mathcal{Q}^\odot(a)\} = \operatorname{sign}_{\mathbb{C}}(a).$$

Proof. For $a = e^{\mathrm{i}\phi}c$, with $c \in \mathbb{R}_+$, a dither $\xi = e^{\mathrm{i}\psi}$, with $\psi \sim \mathcal{U}(0, 2\pi)$, and defining $\alpha := \phi + \psi$, one can show

$$\begin{aligned}
\mathbb{E}_\xi[\mathcal{Q}^\odot(a)] &= \frac{c_\odot}{2\pi}\int_0^{2\pi} \mathcal{Q}(ae^{\mathrm{i}\psi})e^{-\mathrm{i}\psi}d\psi \\
&= \frac{1}{2}e^{\mathrm{i}\,\phi}\int_0^{\frac{\pi}{2}} \mathcal{Q}(e^{\mathrm{i}\,\alpha})e^{-\,\mathrm{i}\,\alpha}d\alpha \\
&= \frac{\sqrt{2}}{2}e^{\mathrm{i}\,\phi+\mathrm{i}\,\frac{\pi}{4}}\int_0^{\frac{\pi}{2}} e^{-\,\mathrm{i}\,\alpha}d\alpha \\
&= e^{\mathrm{i}\phi} = \operatorname{sign}_{\mathbb{C}}(a).
\end{aligned}$$

□

Here lies the big difference with the additive scheme: the multiplicative dither loses information from the signal in expectation. One of its strength, being independent of the amplitude, induces a weaker reconstruction guarantees as part of the information about $\boldsymbol{\Phi x}$ is lost and cannot be retrieved regardless the number of measurements m used.

7.3 Limitations of Phase-Only acquisition

Because the multiplicative dither is linked to the Phase-Only acquisition, we now focus in this section on the limitations that this acquisition process

might encounter with Fourier based measurements. For any algorithms to be able to reconstruct all sparse vectors from their measurements, be it quantized or other acquisition modality, these measurements need to be different from one another. This statement is obvious in the case of classic linear measurements where, if $\boldsymbol{\Phi} \in \mathbb{C}^{m\times N}$ follows the (ℓ_2, ℓ_2)-RIP$(\delta, 2s)$, two different sparse vectors $\boldsymbol{x}_0 \neq \boldsymbol{x}_1 \in \bar{\Sigma}_s^N$ once measured can be said to follow

$$(1-\delta)\|\boldsymbol{x}_0 - \boldsymbol{x}_1\|_2^2 \leq \tfrac{1}{m}\|\boldsymbol{\Phi}\boldsymbol{x}_0 - \boldsymbol{\Phi}\boldsymbol{x}_1\|_2^2 \leq (1+\delta)\|\boldsymbol{x}_0 - \boldsymbol{x}_1\|_2^2,$$

which simply implies that $\boldsymbol{\Phi}\boldsymbol{x}_0 \neq \boldsymbol{\Phi}\boldsymbol{x}_1$.

This simple observation cannot be assumed to be true when using harsh deterministic quantizer (*e.g.*, $\mathcal{Q}_\epsilon(\cdot)$) [Jac+13]. Although the Phase-Only operator can be seen has having an infinite resolution in the phase domain where the other quantizers are limited to a fixed set of values (*e.g.*, number of quadrants), this acquisition can also suffer from ambiguous measurements similar to the one experienced by $\mathcal{Q}_\epsilon$ in Chapter 4 and 5.

For example, one can find 2 vectors $\boldsymbol{x}_0, \boldsymbol{x}_1 \in \mathbb{C}^N$, with $\boldsymbol{x}_0 \neq \boldsymbol{x}_1$, such that

$$\operatorname{sign}_{\mathbb{C}}(\boldsymbol{\Phi}\boldsymbol{x}_0) = \operatorname{sign}_{\mathbb{C}}(\boldsymbol{\Phi}\boldsymbol{x}_1), \tag{7.2}$$

which simply means that $\angle(\boldsymbol{\Phi}\boldsymbol{x}_0) = \angle(\boldsymbol{\Phi}\boldsymbol{x}_1)$. One obvious case is when the two signals are the same but with different amplitude *e.g.*, $\boldsymbol{x}_0 = a\boldsymbol{x}_1$, with $a \in \mathbb{R}_+$. These ambiguous scenarios are, however, inconsequential in the setting of radar estimation as the absolute amplitude of reconstructed radar scene is rarely of interest compared to the relative power between elements of $\boldsymbol{x}$. Other pairs $\boldsymbol{x}_0$, $\boldsymbol{x}_1$ that satisfy (7.2) can be found using the properties of the Fourier transform. For two measured signals to have the same phase, there exists a vector $\boldsymbol{H} \in \mathbb{R}_+^m$, such that

$$\operatorname{sign}_{\mathbb{C}}(\boldsymbol{\Phi}\boldsymbol{x}) = \operatorname{sign}_{\mathbb{C}}(\boldsymbol{H} \odot \boldsymbol{\Phi}\boldsymbol{x}),$$

for a given vector $\boldsymbol{x} \in \mathbb{C}^N$. Given the properties of Fourier transforms, it means that this filter $\boldsymbol{H} \in \mathbb{R}_+^m$ can be expressed from the measurements domain to the signal domain as $\boldsymbol{H} \odot \boldsymbol{\Phi}\boldsymbol{x} = \boldsymbol{\Phi}(\boldsymbol{h} * \boldsymbol{x})$, with $\boldsymbol{h} \in \mathbb{C}^N$ being

the Fourier transform of $\boldsymbol{H}$. In words, any pairs of vectors $\boldsymbol{x}_0, \boldsymbol{x}_1$ will give ambiguous measurements if one vector can be expressed as the convolution of the other with a vector whose Fourier transform is strictly positive.

Fig. 7.4 illustrates this ambiguous scenario for a 1-sparse vector $\boldsymbol{x}$ and a filter $H(t) = 1 + \alpha \cos(2\pi \frac{t}{3T_s})$, with $|\alpha| \leq 1$ and $\frac{1}{T_s}$ the sampling frequency. Because all components of the resulting $\boldsymbol{H}$ are positive, the phase of the measured signals, $\boldsymbol{r}_0 = \boldsymbol{\Phi x}$ and $\boldsymbol{r}_1 = \boldsymbol{H\Phi x}$, are identical. The ambiguous

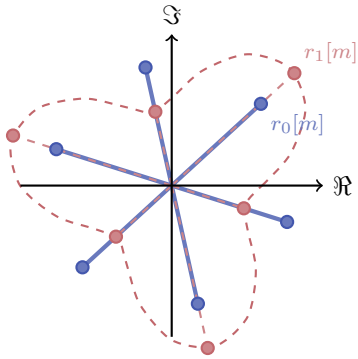


Figure 7.4: Example of an ambiguous scenario where the PO measurements from the **blue** and **red** signals are identical

scenario presented in the measurements domain in Fig. 7.4 is also represented in the signal domain in Fig. 7.5. The convolution between the vector $\boldsymbol{x}$ and $\boldsymbol{h}$ is represented in **red**, where the sparsity of the resulting vector has increased.

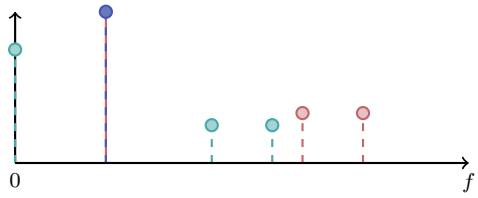


Figure 7.5: Example, in the frequency domain, of an ambiguous scenario where the PO measurements from the $\boldsymbol{x}$ and $\boldsymbol{h} * \boldsymbol{x}$ signals are identical, the filter $\boldsymbol{h}$ is represented in **green**

The mere existence of these ambiguous scenarios for Fourier based model has a dire impact on the maximum theoretical guarantees that the PO ac-

quisition, and consequently the $\mathcal{Q}^{\odot}(\cdot)$ acquisition scheme, can have. Indeed, theoretical guarantees of the type obtained for PO-CS in [JF21; Feu+19] and in Chapter 6 for complex Gaussian matrices are out of reach for Fourier measurements as some vectors fall in these ambiguous scenarios. But this fact should not disqualify completely the PO acquisition, and its quantized counter-part $\mathcal{Q}^{\odot}(\cdot)$, in the context of radar signal processing.

Given the conditions to have sparse vectors whose PO measurements are ambiguous, one can observe that one vector, because of the convolution with $\boldsymbol{h}$, has a smaller support than the other (as depicted in Fig. 7.5). This has several consequences in the context of the estimation of sparse radar scene. One the one hand, if the lower sparsity level signal is the one observed by the radar, then sparsity promoting procedures, *e.g.*, hard-thresholding, will favour this vector over the larger sparsity one during the reconstruction. On the other hand, for a sparse radar scene to be ambiguous with another one once measured is fairly unlikely. Indeed, the power reflected by each target depends on different factors and their phases are often modelled as random [Sko80]. But the factorization of the scene as $\boldsymbol{x}_0 = \boldsymbol{h} * \boldsymbol{x}_1$ requires a specific structure on both the support and the actual complex values of $\boldsymbol{x}_0$ with respect to the individual values of $\boldsymbol{x}_1$ to be valid. All of this makes us confident that using 1-bit sensing with multiplicative dithering can provide a good trade-off between performances and cost of implementation.

7.4 Reconstruction Guarantee

We now study the reconstruction guarantees of the proposed quantization scheme using the PBP algorithm. To that end, we compare the reconstruction of the multiplicatively dithered phase quantizer $\mathcal{Q}^{\odot}(\cdot)$ to the Phase Only acquisition $\operatorname{sign}_{\mathbb{C}}(\cdot)$. Indeed, as indicated by Lemma 7.1, the proposed 1-bit scheme will be lower bounded by the performances of this acquisition scheme.

For a support $\mathcal{S}$ defined as the index set given by the hard thresholding of the back-projection algorithm from multiplicatively dithered 1-bit data,

the reconstruction can be expressed as:

$$\begin{aligned}\|\boldsymbol{x}-\hat{\boldsymbol{x}}\|_2 =& \|\boldsymbol{x}-\tfrac{1}{m}\big(\boldsymbol{\Phi}^H\mathcal{Q}^{\odot}(\boldsymbol{\Phi}\boldsymbol{x})\big)_{\mathcal{S}}\|_2\\ \leq& 2\|\boldsymbol{x}-\tfrac{1}{m}\big(\boldsymbol{\Phi}^H\mathcal{Q}^{\odot}(\boldsymbol{\Phi}\boldsymbol{x})\big)_{\mathcal{T}}\|_2\\ \leq& 2\|\boldsymbol{x}-\tfrac{1}{m}\big(\boldsymbol{\Phi}^H\operatorname{sign}_{\mathbb{C}}(\boldsymbol{\Phi}\boldsymbol{x})\big)_{\mathcal{T}}\|_2\\ &+\frac{2}{m}\|\big(\boldsymbol{\Phi}^H(\operatorname{sign}_{\mathbb{C}}(\boldsymbol{\Phi}\boldsymbol{x})-\mathcal{Q}^{\odot}(\boldsymbol{\Phi}\boldsymbol{x}))\big)_{\mathcal{T}}\|_2 \end{aligned} \tag{7.3}$$

with $\mathcal{T} := \operatorname{supp}(\boldsymbol{x}) \cup \mathcal{S}$. The triangle inequality highlights two terms in (7.3), first the reconstruction using Phase-Only measurements and then a degradation term between the 1-bit measurements and the PO measurements.

A parallel can be drawn between (7.3) and the interplay between the 1-bit acquisition with additive dithering and the classic acquisition without quantization. As the former tends to the latter for high number of measurements [XJ19].

As highlighted in the previous section, the Phase-Only acquisition, using Fourier based measurements, cannot guarantee uniform reconstruction results. Consequently, we only focus on a non-uniform guarantes for the difference between the Phase-Only and 1-bit quantization with multiplicative dithering.

While there is no bound on the reconstruction of PBP using Phase-Only measurement, reconstructing signals from the phase of their Fourier measurements has been a subject of interest for decades. Oppenheim in various publications [OL81; OHL82] tackled this issue for images. They proposed a reconstruction algorithm that can reconstruct images from the phases of their Fourier transform. This algorithm however relies on the fact that considered images are real. More recently, authors in [Feu+19; JF21] studied the recovery guarantees that one can expect from Phase-Only measurement using the framework of Compressive Sensing. Their results applied to complex sparse vectors but assumed a linear model where the measurement matrix follows the (ℓ_1, ℓ_2)-RIP which cannot be directly applied to Fourier based measurements. The reconstruction of complex sparse vector from PO Fourier measurements is still an open and interesting question. However and as will be highlighted in the simulations in Section 7.5, the theoretical limitations of Phase-Only acquisition does not mean

that the reconstruction from these measurements are of poor quality.

Let us now study a non-uniform bound on the ℓ_2-degradation between the PO and 1-bit with multiplicative dithering. For the coming proof, we consider the canonic multiplicative dither whose phase follows a uniform distribution *i.e.*, $\angle\boldsymbol{\xi} \sim \mathcal{U}^m(0, 2\pi)$. The difference between this unstructured random dither and the structured single tone dithering introduced in Sec. 7.2 will be thoroughly studied in the simulations in Section 7.5.

Theorem 7.2. *For a given complex s-sparse vector $\boldsymbol{x} \in \tilde{\Sigma}_s^N$, and a measurement matrix $\boldsymbol{\Phi} \in \mathbb{C}^{m\times N}$ that follows the (ℓ_2, ℓ_2)-RIP$(2s, \delta)$ and with the PO $(\mathrm{sign}_{\mathbb{C}}(\cdot))$ and the 1-bit quantization with multiplicative dithering $(\mathcal{Q}^{\odot}(\cdot))$, and for all support set $\mathcal{T}$ of size $2s$. One can bound, with a probability of failure upper-bounded by* $\exp\left(-\frac{m\gamma^2}{4c_{\odot}^2}\right)$,

$$\|\left(\boldsymbol{\Phi}^T(\mathrm{sign}_{\mathbb{C}}(\boldsymbol{y}) - \mathcal{Q}^{\odot}(\boldsymbol{y}))\right)_{\mathcal{T}}\|_2 < 2m\gamma \tag{7.4}$$

provided $m \geq 8\gamma^{-2}s(\log(\frac{eN}{2s}) + 2(1 + \frac{\pi}{\gamma\sqrt{2}}))$.

The bound in (7.4) combined with the condition imposed on the number of measurement m shows that the discrepancy between the reconstruction using quantized and PO measurements follows $\mathcal{O}(m^{-\frac{1}{2}})$. In other words, the ℓ_2 degradation between the PO and multiplicative dithered measurements can be made arbitrary small given a sufficiently high number of measurements.

7.5 Simulations

We now assess the developed multiplicative scheme by carrying out Monte-Carlo simulations. 100 runs are performed for each set of parameters. At each run, a s-sparse vector is generated. The support is set according to a uniform distribution, the amplitudes of each non zero component are then generated as random uniform variables and given a random phase. The resulting s-sparse vector is then normalized. The linear measurements are generated by the multiplication of these sparse vectors by a matrices made

of elements of a Fourier transform. We define the dimensions of the sparse vector $\boldsymbol{x} \in \mathbb{C}^N$ with $N = 256$ and we vary the number of measurements as $m = \mu N$ with $\mu \in [2^{-6}, 2^4]$. When $\mu \leq 1$ we sub-sample the Fourier transform, which, in the FMCW radar model amounts to sub-sampling the elements of one demodulated chirp (see Chap. 4). When $\mu > 1$ then whole repetitions of the Fourier matrix are taken corresponding thus to multiple consecutive chirps. Five different acquisitions are compared: the classic linear acquisition ( ), the one bit-acquisition with additive dithering $\mathcal{Q}_\epsilon^+$ ( ), the phase only acquisition $\text{sign}_{\mathbb{C}}$ ( ), and 1-bit acquisition with multiplicative dithering $\mathcal{Q}^\odot$ ( ). These 1-bit schemes with dithering are also compared with the deterministic 1-bit acquisition $\mathcal{Q}_\epsilon$ ( ). Those quantized data are then processed with the two different algorithms that were introduced in Chapter 3, namely PBP and QIHT.

7.5.1 ℓ_2 reconstruction

Fig. 7.6 compares the different acquisition schemes using the PBP algorithm. One can first see that the additive dithering behaves as expected by the theory and has a ℓ_2 error that decreases as $\mathcal{O}(m^{-\frac{1}{2}})$ (see Part II and [XJ19]). While this behaviour guarantees an arbitrary low reconstruction error for a number of measurements m high enough, for low number of measurements (*e.g.*, $m \leq 4N$) it is clearly outperformed by the other low resolution schemes. The random multiplicative dithering achieves a reconstruction quality that is better than the additive dithering for any number of measurements presented in Fig. 7.6. Indeed, for $\mu \leq 1$ it follows closely the non-dithered curve ($\mathcal{Q}_\epsilon$) before outperforming it and resolving to the performances of the Phase-Only acquisition. This is consistent with the Lemma 7.1 and the non-uniform bound developed in Theorem 7.2. The only caveat to these performances is that for higher values of m, the additive scheme reaches the performances of the linear classic acquisition without quantization, thus possibly reaching a perfect reconstruction while the multiplicative scheme will still be bounded by the PO performances. It is important to note that, increasing the number of measurements has also a price in terms of data transfer, thus, the quality of the reconstruction of the multiplicative scheme combined with the relative low values of m required for its success makes it an appealing alternative.

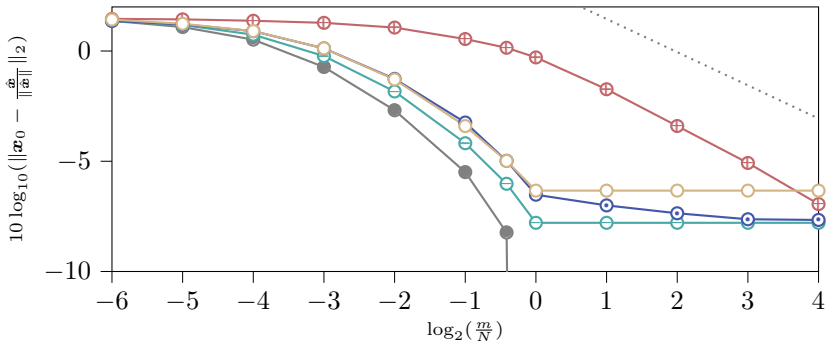


Figure 7.6: Comparison of different reconstructions using PBP between different quantization schemes for $s = 10$, 1-bit with additive dither in **red**; 1-bit without dither in **yellow**; 1-bit with multiplicative dither in **blue**; Phase-Only acquisition in **green**; without quantization in **gray**; the dotted curve in **gray** represent $\mathcal{O}(m^{-\frac{1}{2}})$.

In Fig. 7.7, we now study the performances of an iterative algorithm, namely QIHT. One can observe that the PO acquisition used in conjunction with QIHT algorithm yield high quality reconstructions. Consequently the multiplicative dithering that has its performances bounded by the PO reconstruction, outperforms the additive dither for any number of measurements below 2^4N. Indeed, it follows the deterministic quantization and then continues to exhibits a rate of $\mathcal{O}(m^{-1})$ while the deterministic quantization plateaus at $\mu \geq 1$. It is interesting to see that his rate of $\mathcal{O}(m^{-1})$ has also been observed in other 1-bit setting but with stronger theoretical guarantees [Jac+13]. We clearly see that, while the multiplicative case does not have as strong guarantees as the additive dither, it still has impressive performances and allows good reconstruction while keeping the number of measurements low, especially given the fact that it requires less complex hardware to apply the dither.

The simulations in Fig. 7.6 and Fig. 7.7 showcase the performances of an ideal multiplicative dithering that has a random phase distributed uniformly in the complex domain. In Fig. 7.8, we compare this random dithering with the proposed structured single tone complex exponential defined in (7.1). At each run, the frequency of this complex exponential Δ_f is chosen at random as $\mathcal{U}([-\frac{1}{2T_s}, \frac{1}{2T_s}])$. This known Δ_f is then used to enforce the consistency between the estimated and received radar signal, as only this value is needed

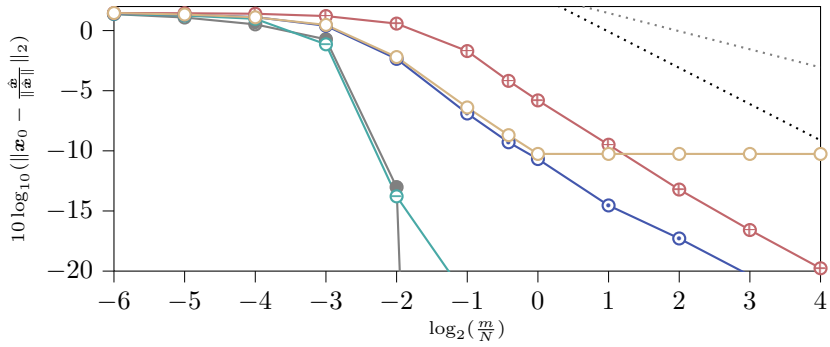


Figure 7.7: Comparison of different reconstructions using QIHT between different quantization schemes for $s = 10$, 1-bit with additive dither in **red**; 1-bit without dither in **yellow**; 1-bit with multiplicative dither in **blue**; Phase-Only acquisition in **green**; without quantization in **gray**; the dotted curve in **gray** represent $\mathcal{O}(m^{-\frac{1}{2}})$, the black dotted curve $\mathcal{O}(m^{-1})$.

to reconstruct the dither used in the quantization. The curves in Fig. 7.8 are almost identical. This shows that the hardware implementation can be further simplified by removing the randomness of the phase and simply using *off-the-shelf* sine and cosine generator at a known frequency. One can now use consistency promoting algorithms without actually measuring or controlling exactly the dither that is applied to the measurements before the quantization.

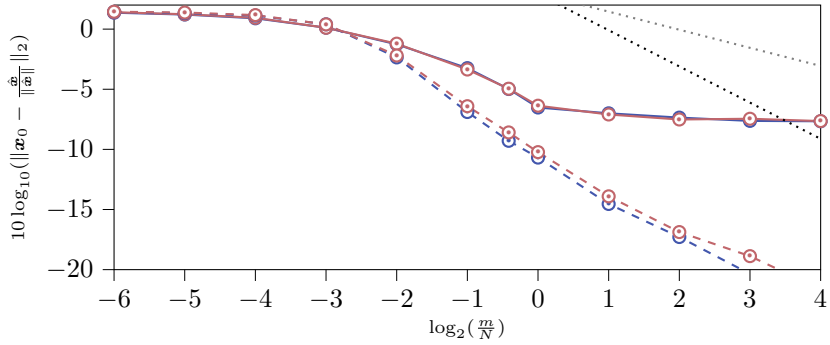


Figure 7.8: Comparison between the random dithering in **blue** and the deterministic and structured dither in **red** for $s = 10$; for the PBP algorithm in solid; and QIHT in dashed; the dotted curve in **gray** represent $\mathcal{O}(m^{-\frac{1}{2}})$, the black dotted curve $\mathcal{O}(m^{-1})$.

In Fig. 7.9, the resilience of the different 1-bit quantization schemes are tested against the imperfect knowledge of the dynamic of the signal $\boldsymbol{\Phi x}$. As mentioned in Section 7.2, an estimate of the signal's dynamic is needed for the additive dithering process and thus impacts its performances, while the multiplicative dither does not depend on the dynamic as it only affects the phase. To mimick the effect of this imperfect knowledge of the signal's dynamic, the size of the additive dither is changed randomly following $\frac{\epsilon}{2} = \|\boldsymbol{\Phi x}\|_\infty 10^\beta$, with $\beta \sim \mathcal{U}(-1, 1)$. One can see that this small

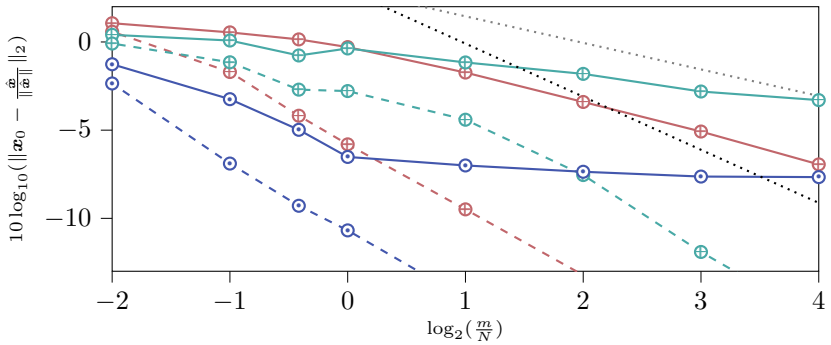


Figure 7.9: Comparison for $s = 10$ of different 1-bit scheme; using, in solid, PBP, QIHT in dashed, 1-bit with additive dither in **red** with perfect dynamic estimation; 1-bit additive dithering with imperfect dynamic estimation in **green**, and the 1-bit with multiplicative dither in **blue**.

unknown on the dynamic induces a deterioration of almost $5dB$ on the additive scheme for both the PBP and QIHT algorithm. This highlights clearly the robustness provided by the multiplicative dither. As only the phase is impacted, the scheme does not require nor suppose anything on the measured signals amplitude, simplifying again the implementation of this scheme in an actual acquisition board.

7.5.2 Support Recovery Performances

The previous simulations focused on the reconstruction error using the ℓ_2-norm. In some application however, recovering the position of the different targets is the only focus, not the actual value of the RCS of the different targets. To that end, one can study the TPR (True Positive Rate), as a way of assessing the support recovery, similarly to Chap. 4. The TPR is defined

as

$$TPR = \frac{|\{i \in [N] \text{ s.t. } |\hat{x}_i| > 0 \text{ and } |x_i| > 0\}|}{K}.$$

In Fig. 7.10 and Fig. 7.11, we compare the TPR for the different algorithms and 1-bit acquisition processes for $s = 4$ and $s = 20$. As seen in [Feu+18b] and in Part II, the additive scheme is not robust to an increase of the sparsity for low number of measurements. For both algorithms, the multiplicative dither, again, either follows the deterministic quantization before outperforming it for higher number of measurements. Using this multiplicative dither, according to the simulations, ensures that the reconstruction cannot be worse than the deterministic quantization and can see up to a $10 - 15\%$ increases in TPR, especially using QIHT in Fig. 7.11.

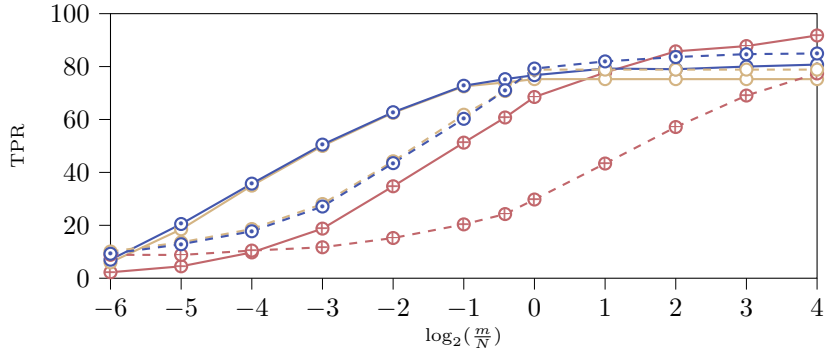


Figure 7.10: Comparison of PBP in terms of TPR for $s = 4$ solid, $s = 20$ in dashed; 1-bit with additive dither in **red**; 1-bit without dither in **yellow**; 1-bit with multiplicative structured dither in **blue**.

7.6 Radar Measurements

To assess the quality of the proposed scheme, we used the dataset featured in [Feu+18b] and in Part II. In this paper, the authors used a 24GHz radar made by RFBEAM [RFB] to measure the signals reflected by 2 target simulators developed by [AMG]. These devices are able to reflect the signal transmitted by the radar back to itself with a specific delay and gain, thus simulating targets at a specific ranges.

A complete explanation of the set-up can be found in [Feu+18b] and in

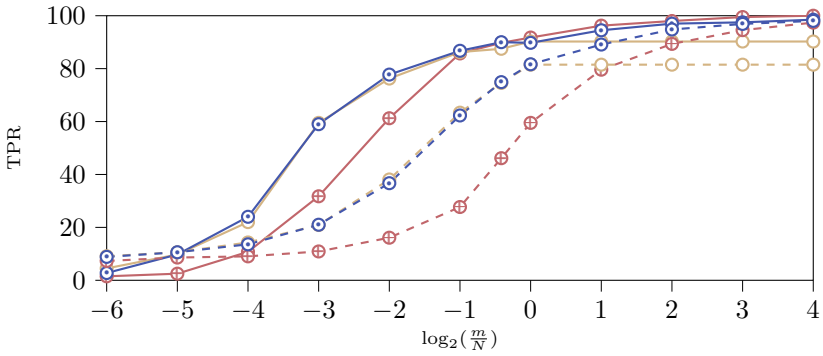
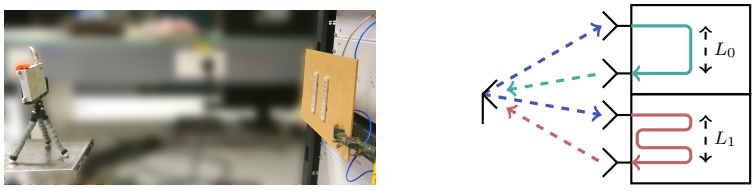


Figure 7.11: Comparison of QIHT in terms of TPR for $s = 4$ solid, $s = 20$ in dashed, 1-bit with additive dither in **red**; 1-bit without dither in **yellow**; 1-bit with multiplicative structured dither in **blue**.



(a) (b)

Figure 7.12: Radar measurements set-up: (a) the KMD2 radar in front of the target simulator; (b) its functional representation.

Chapter 4. To study the effect of a high number of targets, these 2-sparse measurements are combined to generate realistic high sparsity measurements by assuming a simple additive model. For each sparsity level, a 100 Monte-Carlo runs are performed. Where, in order to simulate a s-sparse vector, $\frac{s}{2}$ experiments with disjoint supports are combined.

In Fig. 7.13, we recreated the setting of Fig. 7.6 using the same parameters and reconstruction algorithm *i.e.*, PBP. To estimate the ℓ_2 error, the vector $\boldsymbol{x}$ is defined as the back-projection of the mean of the linear measurements to which an oracle hard-thresholding using the set support by the target simulators is applied. The observations made in the simulations are confirmed using real measurements. There is a stark degradation of the additive scheme for high sparsity level ($s = 10$) with respect to the multiplicative dithering for low number of measurements (*e.g.*, $m \leq 4N$).

Although done in a lab using target simulators, these measurements are not exactly noiseless, as seen by the ℓ_2-reconstruction reached by the classical linear measurements which saturates around $-7dB$, giving thus an approximation of the SNR during the acquisition. This value bounds the rest of the different schemes and also explains why the non-dithered scheme also varies with the number of measurements. Indeed, the noise still acts as an imperfect dither and thus impacts the reconstruction. One can observe, although admittedly small, the beneficial effect of the multiplicative dither on the reconstruction. Similarly to the previous section, in Fig. 7.14, the different

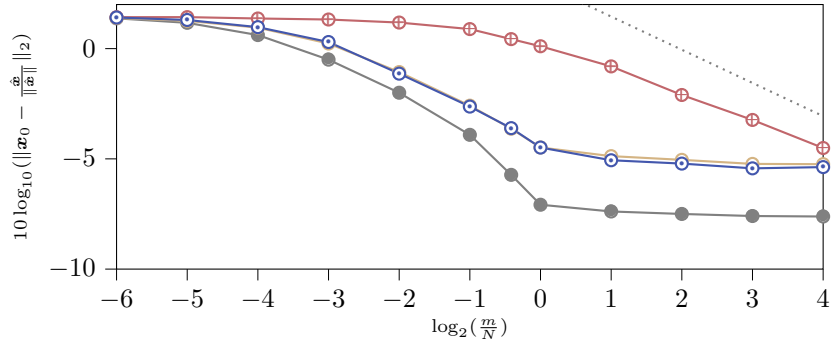


Figure 7.13: ℓ_2 reconstruction for $s = 10$ with PBP using actual radar measurements, 1-bit with additive dither in **red**; 1-bit without dither in **yellow**; 1-bit with multiplicative structured dither in **blue** ; without quantization in **gray**; the dotted curve in **gray** represent $\mathcal{O}(m^{-\frac{1}{2}})$.

acquisitions are compared using the TPR. Again the observations made in Fig. 7.10 are again confirmed here. For low sparsity level, the multiplicative dither outperforms the deterministic quantization. Increasing the sparsity level to $s = 20$, one can also observe a drop in performances of the additive scheme that is not experienced by the multiplicative one that only resolves — as also observed in Fig. 7.13 — to the deterministic 1-bit quantizer.

The results showcased in this section clearly show that using a multiplicative dither provides the best trade-off between hardware requirements and performances between the robustness of the deterministic scheme for high sparsity level and the high quality reconstruction provided by the additive dithering.

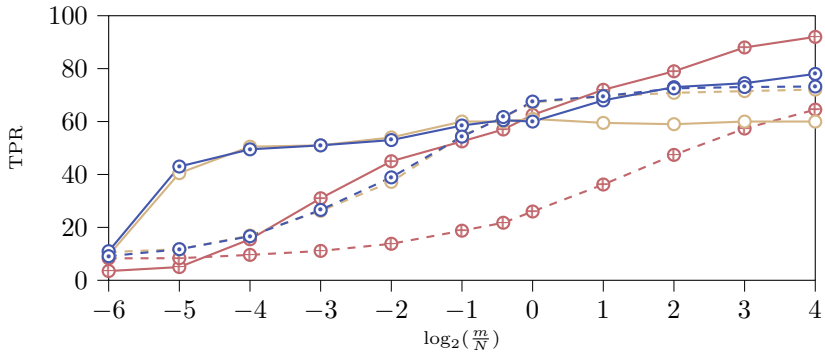


Figure 7.14: Comparison of different TPR using actual radar measurements, for $s = 4$ solid, $s = 20$ in dashed, 1-bit with additive dither in **red**; 1-bit without dither in **yellow**; 1-bit with multiplicative structured dither in **blue**.

7.7 Discussion

In this chapter, we showed that using a multiplicative dithering with radar measurements is a viable trade-off between performances, hardware complexity and theoretical guarantees. Indeed, this work stems from the fact that using an additive dither increases dramatically the complexity of the hardware, negating the gains of the low resolution 1-bit acquisition. The multiplicative dithering of radar signal can be implemented efficiently and relies on already existing hardware architectures. Although the theory uses a dither with random phase, we showed that this condition can be relaxed to a single tone complex exponential, simplifying the architecture even further and relaxing the conditions on reconstruction algorithms that uses the consistency. The developed theory showed that the reconstruction error of this scheme is linked to the one of the Phase-Only acquisition process. Moreover, a non-uniform bound on the discrepancy between these two modalities was also studied and shown to behave as $\mathcal{O}(m^{-\frac{1}{2}})$. Next, the simulations showed that the proposed multiplicative dithering 1-bit scheme provides a good trade-off between the number of measurements needed and the sensitivity to the sparsity level of the signal. These results were then confirmed using real radar measurements.

Future works will focus on the two main aspects of this chapter: the practicality of the proposed scheme and theoretical guarantees and limita-

tions of the PO acquisition. The architecture for the multiplicative dithering presented in Fig. 7.3 will be implemented in a radar prototype and is effectiveness demonstrated. The extension of the model to the angle of arrival will also be studied. This chapter also showed that uniform reconstruction guarantees for Fourier based measurements using the Phase-Only acquisition cannot be developed. This was shown by highlighting that PO measurements can be ambiguous for pairs of sparse vectors. It would be, however, interesting to study this problem further by developing guarantees that can properly deal with these ambiguities.

7.8 Proofs

We first develop a lemma that along with Lemma 7.1 will be used in the main proof.

Lemma 7.3 (Bound related to phase error)**.** *For* $\mathcal{Q}^{\odot}(\cdot)$ *and* $\operatorname{sign}_{\mathbb{C}}(\cdot)$ *for any* $a \in \mathbb{C}$*, one can bound*

$$|\mathcal{Q}^{\odot}(a) - \operatorname{sign}_{\mathbb{C}}(a)| \leq \frac{\pi}{4}.$$

Proof. Let us define, without loss of generality, $a = e^{j\phi}$ and $\xi = e^{j\psi}$, one can first bound

$$\begin{aligned} |\angle\mathcal{Q}^{\odot}(a) - \angle\operatorname{sign}_{\mathbb{C}}(a)| &= |\angle\mathcal{Q}(e^{j\phi+j\psi})e^{-j\psi} - \angle e^{j\phi}| \\ &= |\angle\mathcal{Q}(e^{j\phi'}) - \angle e^{j\phi'}| \leq \frac{\pi}{4}, \end{aligned}$$

with $\phi' := \phi + \psi$.

Considering now a phase difference of β such that $\angle\mathcal{Q}^{\odot}(a) - \angle\operatorname{sign}_{\mathbb{C}}(a) = \beta$, one can bound

$$|\mathcal{Q}^{\odot}(a) - \operatorname{sign}_{\mathbb{C}} a| = |\sqrt{2}\frac{\pi}{4}e^{\mathrm{i}\beta} - 1| = 1 + 2\left(\frac{\pi}{4}\right)^2 - 2\sqrt{2}\frac{\pi}{4}\cos\beta.$$

Setting $\beta = \frac{\pi}{4}$ finally gives the desired bound

$$|\mathcal{Q}^{\odot}(a) - \text{sign}_{\mathbb{C}}(a)| = 1 - (1 - \frac{\pi}{4})\frac{\pi}{2} \leq \frac{\pi}{4}. \tag{7.5}$$

□

Upper bounding (7.5) by $\frac{\pi}{4}$ is made out of convenience and to highlight the fact that $\mathcal{Q}^{\odot}(\cdot)$ effectively quantizes the phase of the signal, not its amplitude.

Lemma 7.3 and Lemma 7.1 combined with *Hoeffding*'s inequality, are used in the following theorem to upper-bound the ℓ_2 discrepancy between the multiplicative and PO reconstruction using PBP.

Theorem 7.2. *For a given complex s-sparse vector $\boldsymbol{x} \in \tilde{\Sigma}_s^N$, and a measurement matrix $\boldsymbol{\Phi} \in \mathbb{C}^{m \times N}$ that follows the (ℓ_2, ℓ_2)-RIP(2s, δ) and with the PO ($\text{sign}_{\mathbb{C}}(\cdot)$) and the 1-bit quantization with multiplicative dithering ($\mathcal{Q}^{\odot}(\cdot)$), and for all support set $\mathcal{T}$ of size $2s$. One can bound, with a probability of failure upper-bounded by $\exp\left(-\frac{m\gamma^2}{4c_{\odot}^2}\right)$,*

$$\|\left(\boldsymbol{\Phi}^T(\text{sign}_{\mathbb{C}}(\boldsymbol{y}) - \mathcal{Q}^{\odot}(\boldsymbol{y}))\right)_{\mathcal{T}}\|_2 < 2m\gamma$$

provided $m \geq 8\gamma^{-2}s(\log(\frac{eN}{2s}) + 2(1 + \frac{\pi}{\gamma\sqrt{2}}))$.

Proof. Starting with

$$\|\left(\boldsymbol{\Phi}^T(\text{sign}_{\mathbb{C}}(\boldsymbol{y}) - \mathcal{Q}^{\odot}(\boldsymbol{y}))\right)_{\mathcal{T}}\|_2 = \sup_{\substack{\boldsymbol{u} \in \mathbb{B}^N \\ \text{supp}(\boldsymbol{u}) \in \mathcal{T}}} \langle \boldsymbol{\Phi}\boldsymbol{u}, \text{sign}_{\mathbb{C}}(\boldsymbol{y}) - \mathcal{Q}^{\odot}(\boldsymbol{y})\rangle. \tag{7.6}$$

Using Lemma 7.1 one can leverage Hoeffding's inequality to upper-bound (7.6). Each quantized i^{th} element of the scalar product are bounded by

$|\boldsymbol{\Phi}_i \boldsymbol{u} \mathcal{Q}^{\odot}(\boldsymbol{y}_i)| \leq |\boldsymbol{\Phi}_i \boldsymbol{u}| c_{\odot} \sqrt{2}$. For a given $\boldsymbol{u}$, (7.6) becomes

$$\mathbb{P}(\|(\boldsymbol{\Phi}^T(\operatorname{sign}_{\mathbb{C}}(\boldsymbol{y}) - \mathcal{Q}^{\odot}(\boldsymbol{y})))_{\mathcal{S}}\|_2 > m\gamma) \leq 2\exp\left(-\frac{m^2\gamma^2}{\|\boldsymbol{\Phi}\boldsymbol{u}\|_2^2 c_{\odot}^2}\right).$$

Leveraging the (ℓ_2,ℓ_2)-RIP$(2s, \delta)$ defined in Def 3.1 we have that $\|\boldsymbol{\Phi}\boldsymbol{u}\|_2^2 \leq m(1+\delta)$, the expression then becomes

$$\begin{aligned}\mathbb{P}\{\|\left(\boldsymbol{\Phi}^T(\operatorname{sign}_{\mathbb{C}}(\boldsymbol{y}) - \mathcal{Q}^{\odot}(\boldsymbol{y}))\right)_{\mathcal{T}}\|_2 > m\gamma\} \\ \leq 2\exp\left(-\frac{m\gamma^2}{2c_{\odot}^2}\right).\end{aligned} \tag{7.7}$$

This relationship is only valid for one given $\boldsymbol{u}$, one needs thus to extend it to all possible $\boldsymbol{u}$ using a covering and a union bound argument. One can define a ρ-covering $\mathcal{J}_\rho$ of the $2s$-sparse space in $\tilde{\Sigma}_{2s}^N$ as defined in Lemma 6.4. Using this covering, one can express any $\boldsymbol{u}$ as $\boldsymbol{u} = \boldsymbol{a} + \boldsymbol{r}$, with $\boldsymbol{a} \in \mathcal{J}_\rho \in \rho\tilde{\Sigma}_{2s}^N$ and $\boldsymbol{r} \in \tilde{\Sigma}_{2s}^N$ with $\|\boldsymbol{r}\|_2 \leq \rho$. The inequality is then extended as follows

$$\begin{aligned}&\|(\boldsymbol{\Phi}^T(\operatorname{sign}_{\mathbb{C}}(\boldsymbol{y}) - \mathcal{Q}^{\odot}(\boldsymbol{y})))_{\mathcal{S}}\|_2 \\ &\qquad = \langle \boldsymbol{\Phi}(\boldsymbol{a}+\boldsymbol{r}), \operatorname{sign}_{\mathbb{C}}(\boldsymbol{y}) - \mathcal{Q}^{\odot}(\boldsymbol{y})\rangle \\ &\qquad \leq m\gamma + \langle \boldsymbol{\Phi}\boldsymbol{r}, \operatorname{sign}_{\mathbb{C}}(\boldsymbol{y}) - \mathcal{Q}^{\odot}(\boldsymbol{y})\rangle,\end{aligned}$$

where (7.7) is extended to all the elements of $\mathcal{J}_\rho$, thanks to a union bound argument. This holds with a probability of failure upper bounded by $2(\frac{eN}{2s})^{2s}(1+\frac{2}{\rho})^{4s}\exp\left(-\frac{m\gamma^2}{c_{\odot}}\right)$ using the bound of $|\mathcal{J}_\rho|$ in Lemma 6.4. One just needs to bound the last term depending on $\boldsymbol{r}$ to conclude the proof. Given its norm and the fact that $\boldsymbol{\Phi}$ follows the (ℓ_2,ℓ_2)-RIP$(2s,\delta)$, one can bound the second term using Lemma 7.3 and *Cauchy-Schwarz* inequality

$$\begin{aligned}\|(\boldsymbol{\Phi}^T(\operatorname{sign}_{\mathbb{C}}(\boldsymbol{y}) - \mathcal{Q}^{\odot}(\boldsymbol{y})))_{\mathcal{S}}\|_2 &\leq m\gamma + \|\boldsymbol{\Phi}\boldsymbol{r}\|_2 \frac{\pi}{4}\sqrt{m} \\ &\leq m\gamma + \sqrt{2}\rho\frac{\pi}{4}m.\end{aligned}$$

Setting $\rho = \frac{4}{\sqrt{2}\pi}\gamma$, finally gives

$$\|(\boldsymbol{\Phi}^T(\operatorname{sign}_{\mathbb{C}}(\boldsymbol{y}) - \mathcal{Q}^{\odot}(\boldsymbol{y})))_{\mathcal{S}}\|_2 \leq 2m\gamma,$$

with a probability of failure upper-bounded by

$$p_\gamma \leq 2(\frac{eN}{2s})^{2s}(1+\frac{\pi}{\gamma\sqrt{2}})^{4s}\exp\left(-\frac{m\gamma^2}{2c_\odot^2}\right). \tag{7.8}$$

Finally, provided that $m \geq 8\gamma^{-2}s(\log(\frac{eN}{2s})+2(1+\frac{\pi}{\gamma\sqrt{2}}))$, one can upper-bound the probability in (7.8) by

$$(\frac{eN}{2s})^{2s}(1+\frac{\pi}{\gamma\sqrt{2}})^{4s}\exp\left(-\frac{m\gamma^2}{2c_\odot^2}\right) \leq \exp\left(-\frac{m\gamma^2}{4c_\odot^2}\right).$$

□

Part IV

Quantizing the Reconstruction

Chapter 8

Binarizing the Reconstruction in 1-bit CS

In this chapter we demonstrate a uniform upper-bound on the ℓ_2 reconstruction of sparse vectors from their (possibly low resolution) measurements using a modified version of the Projected Back-Projection (PBP) algorithm referred to as Quantized PBP (QPBP). Lowering the resolution directly in the processing can yield more cost or power efficient architectures [CA04] as the arithmetical operations used to perform the reconstruction are simplified. More specifically, we study two different types of 1-bit back-projection operators. The first being the direct quantization of the full matrix used for the back-projection, lowering thus its resolution to 1-bit. The application of this modified back-projection is still done through a classic matrix-vector multiplication of known complexity $\mathcal{O}(mN)$. The second quantization process takes advantage of back-projection operators that can be factorized in multiple matrices with fewer non-zero elements than the mN of the full matrix $\mathbf{\Phi}^H$, which results in faster matrix-vector multiplications. The Discrete Fourier transform matrix, for example, can be factorized into $\log_2(N)$ matrices of size $N \times N$ with $2N$ non-zero elements on each sub-matrices. This ubiquitous fast matrix vector multiplication representation of the Fourier transform is known as the Fast Fourier Transform (FFT) [CT65]. In the context of QPBP, each non-zero element of the sub-matrices is quantized in order to benefit both from the simplified

operations and to keep the computation complexity unchanged. For both those quantized reconstruction schemes, we show that the ℓ_2 error obtained using these modified QPBP algorithm decay as $\mathcal{O}(m^{-\frac{1}{2}})$ when applied to both linear and 1-bit measurements. Those theoretical bounds are then confirmed using extensive Monte-Carlo simulations.

8.1 Problem Statement

In order to recover a signal x from its measurements $\boldsymbol{z} = \mathsf{A}(\boldsymbol{\Phi}\boldsymbol{x})$ (*e.g.*, $(\cdot)$ or $\mathcal{Q}_\epsilon^+(\cdot)$), most algorithms use in their estimation process the back-projection of vectors from the measurements domain $\mathbb{C}^m$ to the signal domain $\mathbb{C}^N$. For example the Back-Projection of the measurements $\boldsymbol{z} \in \mathbb{C}^m$ is

$$\tilde{\boldsymbol{x}} = \boldsymbol{\Phi}^H \boldsymbol{z}.$$

Depending on the considered application, applying this back-projection can be seen as the Maximum Likelihood Estimator (MLE) of $\boldsymbol{x}$ given $\boldsymbol{\Phi}$ and $\boldsymbol{z}$ [Kay93]. Adding a hard thresholding H_s on $\tilde{\boldsymbol{x}}$ corresponds to the well known Back-Projection algorithm that has been extensively studied throughout this thesis. But other algorithms such as IHT and its quantized variant QIHT as well as others [FR13; Fou17] depend on the back-projection to estimate this vector $\boldsymbol{x}$.

For these algorithms to be attractive, it is thus capital, to have a Back-Projection that can be computed efficiently. To that end, we study in the first part of this chapter Back-Projection operators $\boldsymbol{\Phi}^H$ that are quantized to 1-bit, *e.g.*, $\boldsymbol{\Psi}^H = \mathcal{Q}_\nu^+(\boldsymbol{\Phi}^H)$, with $\|\boldsymbol{\Phi}\|_{\infty,\infty} \leq \frac{\nu}{2}$.

High resolution back-projection implemented on dedicated hardware, such as Field Progammable Gate Array (FPGA), are usually computed using high resolution multipliers. The back-projection is simply a weighted sum expressed as

$$\tilde{x}_i = \sum_j^N \phi_{ij}^* z_j.$$

These N multiplications between the measurements and the different weights ϕ_{ij}^* must be computed at high resolutions, requiring high cost or power [EL04; Bha+11].

In the 1-bit case, the weights ϕ_{ij}^* are now sent, thanks to the quantizer $\mathcal{Q}_\nu^+$ to 4 points in the complex plane, *e.g.*, $\psi_{ij}^* \in [1, -1, \mathsf{i}, -\mathsf{i}]$.

This means that the estimate $\tilde{x}_i$ which is computed as $\tilde{x}_i = \sum_j^N \psi_{ij}^* z_j$, amounts simply to summing the different measurements z_j with a changed sign or by multiplying by $\pm\mathsf{i}$.

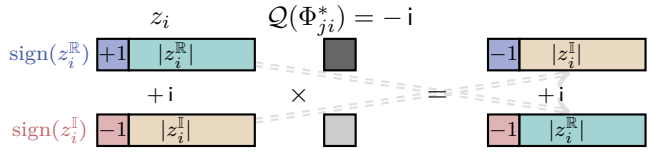


Figure 8.1: Representation of the multiplication of a complex measurement $\boldsymbol{z}_i$, represented in complex binary form, by $\mathcal{Q}(\boldsymbol{\Phi}_{ji}^*) = -\mathsf{i}$.

After the quantization, each complex measurement is represented using 2 arrays of b-bits with 1-bit each for the sign of the real and imaginary part and the $b-1$ bits devoted to the quantized measurements of resolution ϵ (see Fig. 8.1). Taking the negative value of z_i amounts to simply changing the bits that are allocated to the sign of the real and imaginary part while multiplying by $\pm\mathsf{i}$ amounts to simply swapping the real and imaginary part of z_i with the correct signs as shown in Fig. 8.1. Part of the complexity related to the multiplication has been drammatically lowered.

This work is also partly connected to the study of mixed operator in CS [HN10], and on sensing matrix corruption [HS10; PCS11]. In [HN10; HS10] the authors considered performing the reconstruction using a sensing operator deviating from the original one, in this context the mismatch between the two operators is seen as multiplicative noise. The main difference with our work (apart from the use of highly quantized measurements) is that, in [HN10], no assumption is made on the structure of the mismatch which results in quite stringent conditions for the reconstruction of signals. The developed theory in [HN10] only allows for $\approx 5\%$ (in a ℓ_2 sense) discrepancy between the two operators. Whereas our proposed scheme creates a mismatch that is as big (in amplitude) as the maximum component of the matrix, but leverages the added dither to obtain reconstruction error that scales down with an increasing number of measurements. In [PCS11], the authors adapted the Message Passing Approach to solve this challenge. Their studies however only revolved around perturbations that can be linked

to a Gaussian distribution.

The claims of this chapter are the following: in the first part, (*i*) we prove a uniform bound on the degradation between PBP and QPBP applied as a simple matrix vector multiplication, and show that it decays as $O(m^{-\frac{1}{2}})$, provided $\boldsymbol{\Phi}$ follows the RIP and for both linear and 1-bit quantized measurements; (*ii*) we extend this uniform bound to matrices that have efficient matrix-vector multiplication through a factorized model, where we only quantized the non-zero elements of this representation. This model is shown to also decays as $O(m^{-\frac{1}{2}})$ for both linear and 1-bit quantized measurements; (*iii*) we validate these results through Monte-Carlo simulations for Fourier and Gaussian complex matrices, which also highlight the necessity of adding the dither before the quantization.

The chapter is organized as follow: in Sec. 8.2 the signal and pertubed reconstruction models are presented, Sec. 8.3 introduces the different notations that are used in the proofs, Sec. 8.4 studies the direct quantization of the matrix used for the Back-Projection, these results are then extended in Sec 8.5 to matrices that have an efficient matrix-vector representation using a factorized model, all of these guarantees are studied using Monte-Carlo simulations in Sec. 8.6, the details of the different proofs are in Sec. 8.8. Finally, we conclude and propose future works in Sec. 8.7.

8.2 Signal and Quantized Reconstruction Model

We focus on the following linear model

$$\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x},$$

where $\boldsymbol{y} \in \mathbb{C}^m$ are the linear measurements, $\boldsymbol{\Phi} \in \mathbb{C}^{m \times N}$ is the measurement matrix and $\boldsymbol{x} \in \tilde{\Sigma}_s^N$ is an s-sparse vector. As in Part II, we also consider the quantization of $\boldsymbol{y}$ to 1-bit using the 1-bit quantizer with additive dithering ($\mathcal{Q}_\epsilon^+$). We furthermore consider, as is usual in compressive sensing, that the matrix $\boldsymbol{\Phi}$ follows the (ℓ_2, ℓ_2)-RIP$(\delta, 2s)$ defined in Def. 3.1. The last assumption on the model is that the measurements are ℓ_∞-bounded, which is an inherent requirement of 1-bit quantization with an additive dither that is scaled according the dynamic of the linear measurements. Having ℓ_∞-bounded measurements also means that the elements of $\boldsymbol{\Phi}$ are also bounded,

with $\|\boldsymbol{\Phi}\|_{\infty,\infty} \leq \frac{\nu}{2}$.

For the reconstruction, we focus on the Projected Back-Projection algorithm [FR13], namely the signal estimate

$$\hat{\boldsymbol{x}} = \tfrac{1}{m}\mathsf{H}_s(\boldsymbol{A}^H\boldsymbol{b}),$$

where $\mathsf{H}_s(\boldsymbol{u})$ is the hard thresholding operator that zeroes all but the s biggest components of $\boldsymbol{u}$ in amplitude. Depending on the combination of $\boldsymbol{A}$ and $\boldsymbol{b}$, we obtain different schemes. In classical PBP, $\boldsymbol{A} = \boldsymbol{\Phi}$ and $\boldsymbol{b} = \boldsymbol{\Phi x}$ [FR13] with the reconstruction bound presented in Theorem 3.4, and in PBPQ $\boldsymbol{A} = \boldsymbol{\Phi}$ and $\boldsymbol{b} = \mathcal{Q}_\epsilon^+(\boldsymbol{\Phi x})$, where the authors in [XJ19] showed that the reconstruction behaves as $\mathcal{O}(m^{-\frac{1}{2}})$. In this chapter, we first study what happens when we quantize $\boldsymbol{\Phi}$, and set $\boldsymbol{A} = \mathcal{Q}_\nu^+(\boldsymbol{\Phi})$, what we refer as the QPBP and QPBPQ algorithms, with the Q added at the end of QPBP depending on the resolution of the measurements. In the rest of this chapter we consider the 1-bit quantizer defined in (3.5) that sends the measurements to $[\pm 1 \pm \mathsf{i}]$ (up to some constant related to the resolution) which is equivalent to $[1, -1, \mathsf{i}, -\mathsf{i}]$ up to a phase-shift. The conclusion regarding the hardware simplification in Sec. 8.2 still stands.

We refer to the perturbed back-projection operator as $\boldsymbol{\Psi}^H \in \mathbb{C}^{N\times m}$. Our study focuses on Back-Projection operators that are the 1-bit equivalent of the classic back-projection with an added dither, where first, the quantization is applied component wise directly on $\boldsymbol{\Phi}^H$ (as studied in Section. 8.4), and second, where it is applied on each non-zero elements of a factorized matrix as will be introduced in Section 8.5.

To study the reconstruction using a perturbed back-projection operator $\boldsymbol{\Psi}^H$, it is sufficient to study the degradation between the classic reconstruction $\boldsymbol{\Phi}^H$ and the modified one, *i.e.*, for any $\boldsymbol{\Psi}^H$ one can bound

$$\begin{aligned}\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2 &= \|\boldsymbol{x} - \tfrac{1}{m}\big(\boldsymbol{\Psi}^H\boldsymbol{z}\big)_{\mathcal{S}}\|_2 \leq 2\|\boldsymbol{x} - \tfrac{1}{m}\big(\boldsymbol{\Psi}^H\boldsymbol{z}\big)_{\mathcal{T}}\|_2 \\ &\leq 2\|\boldsymbol{x} - \tfrac{1}{m}\left(\boldsymbol{\Phi}^H\boldsymbol{z}\right)_{\mathcal{T}}\|_2 + \tfrac{2}{m}\|\left((\boldsymbol{\Phi}^H - \boldsymbol{\Psi}^H)\boldsymbol{z}\right)_{\mathcal{T}}\|_2 \end{aligned} \quad (8.1)$$

where $\mathcal{S}$ is the support obtained using the hard-thresholding on $\boldsymbol{\Psi}^H\boldsymbol{z}$, $\mathcal{T} := \mathcal{S} \cup \mathrm{supp}(x)$ with $|\mathcal{T}| \leq 2s$ and $\boldsymbol{z} \in \mathbb{C}^m$ are the measurements (quantized or not). Depending on the type of measurement $\boldsymbol{z}$, one can easily upper-bound

the first term of (8.1). If $\boldsymbol{z} = \boldsymbol{\Phi}\boldsymbol{x}$, then the first terms is upper-bounded by the reconstruction of PBP for linear measurements as in Theorem 3.4. In that case, the first term decays as $\mathcal{O}(m^{-\frac{1}{2}})$. For quantized measurements, the first term is simply bounded by $\mathcal{O}((1+\epsilon)m^{-\frac{1}{2}})$, as shown in [XJ19] as it corresponds to PBPQ (see Thm. 3.3). Only the degradation between the PBP(Q) and QPBP(Q) remains to be bounded. This is the focus of the next sections.

8.3 Notations

The operator $\mathrm{diag}_k(\boldsymbol{A})$ creates the block diagonal matrix consisting of k repetition of the square matrix $\boldsymbol{A} \in \mathbb{C}^{d\times d}$, so $\mathrm{diag}_k(\boldsymbol{A}) \in \mathbb{C}^{kd\times kd}$. The identity matrix is written as $\boldsymbol{I}_d = \mathrm{diag}_d(1) \in \mathbb{R}^{d\times d}$. $\mathrm{Tr}(\cdot)$ is the trace operator such that $\mathrm{Tr}(A) = \sum_i^d A_{i,i}$, $\|\boldsymbol{A}\|$ is the spectral norm of $\boldsymbol{A} \in \mathbb{C}^{m\times N}$. $\|\boldsymbol{A}\|_2 = (\sum_{i,j} |\boldsymbol{A}_{i,j}|^2)^{\frac{1}{2}}$ is the Frobenius norm. For a matrix $\boldsymbol{A}$, we define the *Schatten* p-norm written as $\|\boldsymbol{A}\|_p$ as applying the ℓ_p-norm on the singular value of $\boldsymbol{A}$, as such $\|\boldsymbol{A}\|_\infty = \|\boldsymbol{A}\|$ is the spectral norm of $\boldsymbol{A}$, $\|\boldsymbol{A}\|_2$ is the Frobenius norm of $\boldsymbol{A}$ and $|||\boldsymbol{A}|||_{p,q} := (\mathbb{E}\{\|\boldsymbol{A}\|_p^q\})^{\frac{1}{q}}$.

8.4 Matrix Multiplication

Let us now study the case where $\boldsymbol{\Psi}^H = \mathcal{Q}_\nu^+(\boldsymbol{\Phi}^H) \in \mathbb{C}^{N\times m}$, *i.e.*, where the back-projection operator is quantized directly to 1-bit with an appropriately scaled dither. Each component of the back-projection matrix $\boldsymbol{\Phi}^H$ is dithered with an independent random uniform variable that is also independent from the possible dither used for quantizing the measurements. We will first study a non uniform upper-bound for a fixed measurement vector $\boldsymbol{z}$ which is ℓ_∞-bounded by $\frac{\epsilon}{2}$. This allows us to cover both classic linear and low resolution measurements by upper-bounding them by $\frac{\epsilon}{\sqrt{2}}$. We then extend it to all s-sparse vectors with and without quantization through a union bound and covering argument.

Before dwelving into the different theorems, let us study a toy example that highlights the gain provided by the addition of a dither to the quantized back-projection.

Considering $\boldsymbol{\Psi}^H = \mathcal{Q}_\nu(\boldsymbol{\Phi}^H)$, *i.e.*, the direct one-bit equivalent of the

measurement matrix without any dither, with $\|\boldsymbol{\Phi}\|_{\infty,\infty} \leq \frac{\nu}{2}$. In that case and with no further assumption on $\boldsymbol{\Phi}$ and $\boldsymbol{z}$, the second term in (8.1) can be bounded by

$$\frac{2}{m}\| \left((\boldsymbol{\Phi}^H - \boldsymbol{\Psi}^H)\boldsymbol{z}\right)_{\mathcal{T}} \|_2 \leq \frac{2\sqrt{2s}}{m} \max_i |\langle (\boldsymbol{\Phi}^H - \boldsymbol{\Psi}^H)_i, \boldsymbol{z}\rangle|. \tag{8.2}$$

Because of the deterministic 1-bit quantization of resolution ν, one can bound $|\Phi^*_{ij} - \Psi^*_{ij}| \leq \frac{\nu}{\sqrt{2}}$. Combining this bound with the ℓ_∞-bound on $\boldsymbol{z}$, the scalar product in (8.2) finally gives

$$\frac{2}{m}\| \left((\boldsymbol{\Phi}^H - \boldsymbol{\Psi}^H)\boldsymbol{z}\right)_{\mathcal{T}} \|_2 \leq 2\sqrt{s}\epsilon\nu. \tag{8.3}$$

The bound in (8.3) clearly highlights the multiplicative nature of the pertubation $\boldsymbol{\Psi}$ from $\boldsymbol{\Phi}$ [HN10; HS10]. If this direct quantization approach is used, then the ℓ_2-reconstruction remains constant regardless of the number of measurements m used in the acquisition. It is possible that some linear systems with specific measurements matrix might be able to exhibit better reconstruction guarantees using this deterministic quantization on the back-projection. In this work however, we target a more general reconstruction bound that can be applied to all systems provided that they follow the RIP property and that the measurements are a ℓ_∞-bounded. To that end, we depart from the simple 1-bit quantization of the back-projection $\mathcal{Q}_\nu(\boldsymbol{\Phi}^H)$ to its dithered counterpart $\mathcal{Q}^+_\nu(\boldsymbol{\Phi}^H)$. Where the fact that $\mathbb{E}\{\mathcal{Q}^+_\nu(\boldsymbol{\Phi}^H)\} = \boldsymbol{\Phi}^H$ will be leveraged to obtain tighter bounds than the one in (8.3).

Let us first look at a non-uniform result on the degration between PBP(Q) and QPBP(Q):

Lemma 8.1. *For a measurement vector $\boldsymbol{z} \in \mathbb{C}^m$, with $\|\boldsymbol{z}\|_\infty \leq \frac{\epsilon}{\sqrt{2}}$, a measurement matrix $\boldsymbol{\Phi} \in \mathbb{C}^{m\times N}$ that follows the (ℓ_2,ℓ_2)-RIP$(\delta, 2s)$ that is, furthermore, ℓ_∞-bounded $\|\boldsymbol{\Phi}\|_{\infty,\infty} \leq \frac{\nu}{2}$, for any support set $\mathcal{T}$ of size $2s$, one can upper-bound the PBP degradation using a modified back-projection operator defined as $\boldsymbol{\Psi}^H := \mathcal{Q}^+_\nu(\boldsymbol{\Phi}^H)$ by*

$$\|((\boldsymbol{\Psi} - \boldsymbol{\Phi})^H\boldsymbol{z})_{\mathcal{T}}\|_2 \leq m\gamma,$$

with a probability exceeding $1 - 2N \exp\big(\frac{-m\gamma^2}{16s\epsilon^2\nu^2}\big)$.

Proof. For a fixed measurement vector $\boldsymbol{z} \in \mathbb{C}^m$, let us start by studying one element $i \in \mathcal{T}$ that makes up the desired ℓ_2-norm. Given that $\mathbb{E}_{\boldsymbol{\Psi}}\{\boldsymbol{\Psi}_i^H \boldsymbol{z}\} = \boldsymbol{\Phi}_i^H \boldsymbol{z}$, and using Hoeffding's lemma [Ver18, Theorem 2.2.6], one can bound

$$|\boldsymbol{\Psi}_i^H \boldsymbol{z} - \boldsymbol{\Phi}_i^H \boldsymbol{z}| > \frac{m\gamma}{\sqrt{2s}} \tag{8.4}$$

with a probability upper-bounded by $2\exp\big(-\frac{m\gamma^2}{16s\epsilon^2\nu^2}\big)$, where we upper-bounded $\max_{j\in[N]} |(\boldsymbol{\Psi}_{ij}^H - \boldsymbol{\Phi}_{ij}^H)\boldsymbol{z}_j| \leq \epsilon\nu$.

Extending this to all elements of the ℓ_2-norm using a union bound on all $i \in [N] \supset \mathcal{T}$ finally gives us :

$$\mathbb{P}\big(\|(\boldsymbol{\Psi}^T\boldsymbol{z} - \boldsymbol{\Phi}^T\boldsymbol{z})_{\mathcal{T}}\|_2 > m\gamma\big) \leq 2N \exp\big(\frac{-m\gamma^2}{16s\epsilon^2\nu^2}\big).$$

□

Lemma 8.1 applies to any support sets $\mathcal{T}$, such that $|\mathcal{T}| \leq 2s$, as (8.4) is extended to $\forall i \in [N]$, which covers all possible supports $\mathcal{T}$. This lemma can now be extended to all s-sparse vectors. This extension, through a covering and union bound argument, will be done for both types of measurements separately, *i.e.*, classic linear measurements and 1-bit ones with additive dithering. Let us start with the linear measurements.

Theorem 8.2. *For all s-sparse vector $\boldsymbol{x} \in \tilde{\Sigma}_s^N$, a measurement matrix $\boldsymbol{\Phi} \in \mathbb{C}^{m\times N}$ that follows the (ℓ_2, ℓ_2)-RIP$(\delta, 2s)$, considering furthermore that $\|\boldsymbol{\Phi}\boldsymbol{x}\|_\infty \leq \frac{\epsilon}{2}$ and that $\|\boldsymbol{\Phi}\|_{\infty,\infty} \leq \frac{\nu}{2}$, for all support sets $\mathcal{T}$ of size $2s$ and using a modified back-projection operator defined as $\boldsymbol{\Psi}^H := \mathcal{Q}_\nu^+(\boldsymbol{\Phi}^H)$, one can upper-bound, with a probability exceeding $1 - 2N \exp\big(\frac{-m\gamma^2}{32s\epsilon^2\nu^2}\big)$,*

$$\|((\boldsymbol{\Psi} - \boldsymbol{\Phi})^H \boldsymbol{\Phi}\boldsymbol{x})_{\mathcal{T}}\|_2 \leq 3m\gamma,$$

provided $m \geq 16s^2\epsilon^2\nu^2\gamma^{-2}\big(\log(\frac{eN}{s}) + 2\log(1 + \frac{2\sqrt{2s}\nu}{\gamma})\big)$.

Theorem 8.2 shows that the discrepancy term will behave as $\mathcal{O}(\frac{s}{\sqrt{m}})$ while the classic PBP estimate follows $\mathcal{O}(\sqrt{\frac{s}{m}})$ (see Rem. 3.3). We conjecture that this mismatch with respect to the sparsity level is an artefact of the proof and can be tightened to $\mathcal{O}(\sqrt{\frac{s}{m}})$. The steps in the proof where, for some vector $\boldsymbol{w} \in \mathbb{C}^N$, we use the bound $\|(\boldsymbol{w})_{\mathcal{T}}\|_2 \leq \sqrt{|\mathcal{T}|}\|\boldsymbol{w}\|_\infty$ could be refined in order to avoid this extra $\sqrt{s}$. This improved proof would then match the behaviour of the bound on classic high resolution PBP estimate.

Lemma 8.1 can also be extended to the case of quantized measurements. The discrepancy between PBPQ and QPBPQ can be upper-bounded using the following theorem.

Theorem 8.3. *For all s-sparse vector $\boldsymbol{x} \in \tilde{\Sigma}_s^N$, a measurement matrix $\boldsymbol{\Phi} \in \mathbb{C}^{m\times N}$ that follows the (ℓ_2, ℓ_2)-RIP$(\delta, 2s)$, considering furthermore that $\|\boldsymbol{\Phi}\boldsymbol{x}\|_\infty \leq \frac{\epsilon}{2}$ and that $\|\boldsymbol{\Phi}\|_{\infty,\infty} \leq \frac{\nu}{2}$, for all support set $\mathcal{T}$ of size $2s$ and using a modified back-projection operator defined as $\boldsymbol{\Psi}^H := \mathcal{Q}_\nu^+(\boldsymbol{\Phi}^H)$, one can upper-bound with a probability exceeding $1 - 2N\exp\left(\frac{-m\gamma^2}{32s\epsilon^2\nu^2}\right)$,*

$$\|((\boldsymbol{\Psi} - \boldsymbol{\Phi})^H \mathcal{Q}_\lambda^+(\boldsymbol{\Phi}\boldsymbol{x}))_{\mathcal{T}}\|_2 \leq 3m\gamma,$$

provided $m \geq 32\epsilon^2\nu^2\gamma^{-2}s^2\left(\log(\frac{eN}{s}) + 2\log(1 + 24\frac{\epsilon\nu^2\sqrt{2m}}{\gamma^2})\right)$.

Theorem 8.3 shows that the upper-bound on the discrepancy between PBPQ and QPBPQ will also follow $\mathcal{O}(sm^{-\frac{1}{2}})$. This behaviour is exactly the same as the one obtained in Theorem 8.2 for the linear measurements. As the proof of Theorem 8.3 follows the same steps as its high resolution counter-part, the bound exhibits the same linear dependency for the sparsity with respect to the number of measurement m. We also conjecture that this bound could be refined to $\mathcal{O}(\sqrt{\frac{s}{m}})$.

The previous two theorems show that directly quantizing the back-projection to 1-bit to be used for the reconstruction yields performances that are similar to their high-resolution counterparts as their reconstructions all follow, up to a multiplicative constant, $\mathcal{O}(m^{-\frac{1}{2}})$. The switch from a high to a low resolution BP only requires an increase of the number of measurements to compensate for this constant (see the simultions results in

Sec.8.6).

8.5 1-bit Factorizable Back-Projection

The previous section showed that it is possible to dramatically lower the resolution of the back-projection operator and still be able to upper-bound the reconstruction obtained by this modified PBP and for it to decay as $\mathcal{O}(m^{-\frac{1}{2}})$. This quantization can provide interesting gains in the hardware implementation but applying the quantized back-projection as a simple matrix multiplication still has a known complexity of $\mathcal{O}(mN)$. However, there exist matrices whose product with another vector can be computed more efficiently. One example is the ubiquitous Fast Fourier Transform [CT65] which has a complexity of $\mathcal{O}(m \log_2(N))$.

In this section, we focus on matrices that exhibit a structure similar to the matrix representation of the FFT, *i.e.*, matrices that can be factorized into a multiplication of sub-matrices that have few non-zero coefficients each. For example, when $m = N$, the Fourier transform can be factorized in $\log_2(N)$ sub-matrices of dimensions $\mathbb{C}^{N\times N}$. Following the Cooley-Tukey (or radix-2) algorithm [CT65], each line has only two non-zero elements. This gives only $N \log_2(N)$ operations, or $2N \log_2(N)$ non-zero elements across the $\log_2(N)$ matrices compared to N^2 in the classic matrix multiplication case.

The factorization of the matrix $\mathbf{\Phi}^H \in \mathbb{C}^{N\times N}$ used for the back-projection can be represented as :

$$\mathbf{\Phi}^H = \prod_{i=1}^{J} \mathbf{\Upsilon}^i \tag{8.5}$$

where J is the number of sub-matrices (*e.g.*, for Fourier transform $J = \log_2(N)$). $\mathbf{\Upsilon}^i \in \mathbb{C}^{N\times N}$ is the $i^{th} \in [J]$ matrix of the factorized model, with an implicit ordering of the form $\prod_{i=1}^{J} \mathbf{\Upsilon}^i = \mathbf{\Upsilon}^J \mathbf{\Upsilon}^{J-1} ... \mathbf{\Upsilon}^1$. We consider sub-matrices $\mathbf{\Upsilon}^i$ with only κ non-zero elements on each rows. So $\|\mathbf{\Upsilon}^i_j\|_0 \leq \kappa$ for $j \in [N]$ and $\|\mathbf{\Upsilon}^i\|_{0,1} \leq \kappa N$.

In this section, we are interested in studying the reconstruction performances of this factorized scheme when the non-zero coefficients of the

matrices $\boldsymbol{\Upsilon}^i$ are quantized to 1-bit. The system can then be represented as

$$\boldsymbol{\Psi}^H = \prod_{i=1}^{J} \mathcal{Q}^+_{\nu_i}(\boldsymbol{\Upsilon}^i), \tag{8.6}$$

where $\mathcal{Q}^+_{\nu_i}(\cdot)$ only quantizes the non-zero elements of each sub-matrix in order to keep the structure and thus the complexity of the computation. The constant ν represents the quantization step size for all sub-matrices, *i.e.*, $\max_{i\in[J]} \|\boldsymbol{\Upsilon}^i\|_{\infty,\infty} \leq \frac{\nu}{2}$. Furthermore, the subscript i on $\mathcal{Q}^+_{\nu_i}(\cdot)$ signifies explicitly that each sub-matrices and its associated κNJ non-zero elements are dithered with different and thus independent random variables following a uniform distribution $\mathcal{U}(-\frac{\nu}{2}, \frac{\nu}{2}) + \mathrm{i}\mathcal{U}(-\frac{\nu}{2}, \frac{\nu}{2})$. This work tackles an issue similar to [CA04], but goes further by developing a reconstruction bound whereas [CA04] only relies on the expected value of the quantized operator.

The model in (8.6) is only valid for $m = N$, we now introduce how this model can be extended to a case where $m > N$. This is motivated by recent publications in the domain of Quantized Compressive Sensing where the authors showed that although the quantity of data was overall reduced, *i.e.*, its bit-rate defined by $\mathcal{B} = m \times b$ (with b being the resolution of the ADCs), inducing repetitions and thus oversampling of the data can be beneficial for extremely quantized scebarios [Feu+18a; Feu+18b; Feu+19; Bou+15]. We focus on physical models where taking more than $m = N$ measurements amounts to sampling repetitions of the signal of interests. This is the case, for example, in FMCW radars where one can sample multiple consecutive received chirps (see Chap. 2).

In order to represent the signal model for $m > N$, we introduce the effect of the added repetition on the possible random sub-sampling of each repetitions from $\boldsymbol{\Phi}\boldsymbol{x}$ by expressing the number of measurement as $m = \rho\mu N$ with $\rho \in \mathbb{N}_+$ being the number of repetitions and $0 < \mu \leq 1$ being the rate of sub-sampling. The Quantized Back-Projection model can then be expressed as

$$\tilde{\boldsymbol{x}} = \boldsymbol{P}^\rho \prod_{i=1}^{J} \mathcal{Q}^+_{\nu_i}(\mathrm{diag}_\rho(\boldsymbol{\Upsilon}^i))\tilde{\boldsymbol{S}}\tilde{\boldsymbol{z}} \tag{8.7}$$

with $\boldsymbol{P}_\rho := (\boldsymbol{I}_N, \ldots, \boldsymbol{I}_N) \in \mathbb{C}^{N\times\rho N}$ is the matrix that sums the ρ repetitions of the estimated back-projection. The sub-sampling matrix $\tilde{\boldsymbol{S}} \in \mathbb{R}^{\rho N\times\rho N}$

can be expressed as

$$\tilde{\boldsymbol{S}} = \begin{pmatrix} \boldsymbol{S}^1 & & 0 \\ & \ddots & \\ 0 & & \boldsymbol{S}^\rho \end{pmatrix}, \tag{8.8}$$

with $\boldsymbol{S}^r \in \mathbb{C}^{N\times N}$ being the sub-sampling of each repetition for $r \in [\rho]$, such that $\sum_r^\rho \mathrm{Tr}(\boldsymbol{S}^r) = m = N\mu\rho$. The model in (8.7) amounts to performing ρ acquisitions of μN measurements and performing ρ Quantized-BP. $\tilde{\boldsymbol{z}} \in \mathbb{C}^{\rho N}$ are the ρ repetitions of the (possibly quantized) measurements.

8.5.1 Reconstruction Bound

Results in Theorems 8.2 and 8.3 are based on *Hoeffding*'s inequality [Ver18, Theorem 2.2.6] that revolves around the fact that each element of the results of the Back-Projection is made of a sum of m independent random variables. Bounding the reconstruction obtained with the quantized and factorized BP cannot yield satisfactory bounds using the same methodology. Indeed, although in model (8.6) the κNJ different coefficients are each quantized with random and independent dithers, each component of the estimated vector $\tilde{\boldsymbol{x}}$ is only made up of κ random variables

$$\tilde{x}_i = \mathcal{Q}_{\nu_J}(\Upsilon_i^J) \prod_j^{J-1} \mathcal{Q}_{\nu_j}(\boldsymbol{\Upsilon}^j)\boldsymbol{z} = \sum_{k\in\mathcal{K}} \mathcal{Q}_{\nu_J}(\Upsilon_{i,k}^J)(\prod_j^{J-1} \mathcal{Q}_{\nu_j}(\boldsymbol{\Upsilon}^j)\boldsymbol{z})_k,$$

with $i \in [N]$ and with $\mathcal{K}$ being the support of Υ_i^J with $|\mathcal{K}| = \|\Upsilon_i^J\|_0 = \kappa$. This severely limits the performances of the bound that one can obtain using *Hoeffding*'s inequality, as it entirely ignores the dimension N of the problem and also any possible sub-sampling as well as the fact that the model is based on the multiplication of multiple matrices. The bound would only leverage the over-sampling factor ρ and as it extends this sum of κ independent variables to $\rho\kappa$.

The problem of finding a tight bound on the reconstruction based on the estimate obtained by the modified back-projection defined in (8.7) must fully leverage the factorized structured and the independence of the different quantized coefficients.

Recent advances in concentration of measures have shown that one can bound the spectral norm of the multiplication of random matrices. More specifically, the authors in [Hua+20] showed that for a independent sequence $\{\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_J\} \subset \mathbb{C}^{N\times N}$ of random matrices, with $\|\boldsymbol{Y}^i\|_2 \leq b_i$ and $|||\boldsymbol{Y}^i - \mathbb{E}[\boldsymbol{Y}^i]|||_{p,2} \leq \sigma_i b_i$, one can prove that

Theorem 8.4 (Remark 5.7 (Uniform Bounds on Factors) from [Hua+20])**.** *Let $v = \sum_i^J \sigma_i^2$ and $B = \prod_i^J b_i$. Then we have an unconditional variant of the concentration bound:*

$$\mathbb{P}\{\|\prod_i^J \boldsymbol{Y}^i - \mathbb{E}[\prod_i^J \boldsymbol{Y}^i]\| \geq tB\} \leq \max(N, e)\exp\left(-\frac{t^2}{2ev}\right)$$

for all $t > 0$.

The parallel between Theorem 8.4 and the search for an upper-bound on the difference between the classical factorized BP in (8.5) and its quantized and randomly dithered counterpart presented in (8.6) is clear.

In order to leverage the results of Theorem 8.4 and to apply it to the model introduced in (8.7), we need the following quantities

Lemma 8.5. *Considering a back-projection model $\boldsymbol{\Phi}^H := \prod_i^J \boldsymbol{\Upsilon}^i$ with $\max_i \|\boldsymbol{\Upsilon}^i\|_{\infty,\infty} \leq \frac{\nu}{2}$ where each $\boldsymbol{\Upsilon}^i \in \mathbb{C}^{N\times N}$ is κ-line sparse,* i.e.*, $\forall i \in [J], \|\boldsymbol{\Upsilon}_j^i\|_0 \leq \kappa$. Using the 1-bit dithered quantizer $\mathcal{Q}_\nu^+(\cdot)$, one can then show that :*

$$\|\mathcal{Q}_{\nu_i}^+(\boldsymbol{\Upsilon}^i)\| \leq \nu\sqrt{\frac{N\kappa}{2}}, \qquad |||\mathcal{Q}_{\nu_i}^+(\boldsymbol{\Upsilon}^i) - \boldsymbol{\Upsilon}^i|||_{2,2}^2 \leq \kappa N\frac{\nu^2}{8}.$$

Proof. Let us start with the spectral norm of $\mathcal{Q}_\nu^+(\boldsymbol{\Upsilon}^i)$. Given the κ-sparse structure of the matrices and the fact that the matrices are quantized at a

resolution of 1-bit, we can easily show that :

$$\begin{aligned}\|\mathcal{Q}^+_{\nu_i}(\boldsymbol{\Upsilon}^i)\| &= \max_{\|\boldsymbol{u}\|=1} \|\mathcal{Q}^+_{\nu}(\boldsymbol{\Upsilon}^i)\boldsymbol{u}\|_2 \\ &\leq \frac{\nu}{\sqrt{2}} \max_{\|\boldsymbol{u}\|=1} \sqrt{\sum_n^N \|\boldsymbol{u}_{\mathcal{S}^i_n}\|_1^2}, \end{aligned} \tag{8.9}$$

where $\mathcal{S}^i_n := \text{supp}(\boldsymbol{\Upsilon}^i_n)$, with $n \in [N]$, is the support set of the n^{th} row of $\boldsymbol{\Upsilon}^i$. As such $|\mathcal{S}^i_n| \leq \kappa$. Here, one can use the pessimistic upper-bound on (8.9), that assumes that all support set $\mathcal{S}_n$ are identical for all $i \in [J]$ and all $n \in [N]$, which in turns means that the vector $\boldsymbol{u}$ that maximizes this expression is $u_i = \frac{1}{\sqrt{\kappa}}$ for $i \in \mathcal{S}^i_n$, 0 elsewhere. This approximation effectively assumes that the submatrices $\boldsymbol{\Upsilon}^i$ have no particular structure in their supports. (8.9) then becomes :

$$\|\mathcal{Q}^+_{\nu}(\boldsymbol{\Upsilon}^i)\| \leq \nu\sqrt{\frac{N\kappa}{2}}.$$

Dealing now with the second term, given the definition of the *Schatten p-norm* with $p = 2$, one can bound:

$$|||\mathcal{Q}^+_{\nu_i}(\boldsymbol{\Upsilon}^i) - \boldsymbol{\Upsilon}^i|||^2_{2,2} = \mathbb{E}\{\|\mathcal{Q}^+_{\nu_i}(\boldsymbol{\Upsilon}^i) - \boldsymbol{\Upsilon}^i\|_2^2\} \leq \kappa N \mathbb{E}\{\|\mathcal{Q}^+_{\nu_i}(\boldsymbol{\Upsilon}^i) - \boldsymbol{\Upsilon}^i\|^2_{\infty,\infty}\}$$

$$\leq \kappa N \max_{j,k} \text{var}(\mathcal{Q}^+_{\nu_i}(\boldsymbol{\Upsilon}^i)_{j,k}).$$

Given that $\mathcal{Q}^+_{\nu_i}(\cdot) \in [\pm\frac{\nu}{2} \pm \frac{\nu}{2}\,\mathrm{i}]$, the variance can be upper-bounded by

$$\text{var}(\mathcal{Q}^+_{\nu_i}(\boldsymbol{\Upsilon}^i)_{j,k}) \leq \frac{\nu^2}{8},$$

by upper-bounding the variance of 1-bit quantized factor of $\boldsymbol{\Upsilon}^i$ of resolution ν by Bernoulli random variables. Which finally gives :

$$|||\mathcal{Q}^+_{\nu_i}(\boldsymbol{\Upsilon}^i) - \boldsymbol{\Upsilon}^i|||^2_{2,2} \leq \kappa N \frac{\nu^2}{8}.$$

□

We will also need the following quantity

Lemma 8.6. *Given J matrices $\boldsymbol{\Upsilon}^i$, with $i \in [J]$ and $\|\boldsymbol{\Upsilon}^i\|_{\infty,\infty} \leq \frac{\nu}{2}$, whose lines are κ-sparse. One can bound,*

$$\|\prod_{i=1}^{J} \mathcal{Q}_{\nu_i}^{+}(\boldsymbol{\Upsilon}^i)\|_{\infty,\infty} \leq \kappa^{-1}\big(\kappa\frac{\nu}{\sqrt{2}}\big)^J.$$

Proof. Leveraging the κ-sparse structure of each lines of the submatrices $\boldsymbol{\Upsilon}^i$, one can bound for a unitary vector $\mathbf{1}_d$ and for a given $i \in [J]$

$$\|\mathcal{Q}_{\nu_i}^{+}(\boldsymbol{\Upsilon}^i)\mathbf{1}_N\|_\infty \leq \max_l \|\mathcal{Q}_{\nu_i}^{+}(\boldsymbol{\Upsilon}^i)_l\|_1 \leq \kappa\frac{\nu}{\sqrt{2}}$$

We can now re-express the desired bound as :

$$\begin{aligned}
\|\prod_{i=1}^{J} \mathcal{Q}_{\nu_i}^{+}(\boldsymbol{\Upsilon}^i)\|_{\infty,\infty} &= \|\big(\prod_{i=2}^{J} \mathcal{Q}_{\nu_i}^{+}(\boldsymbol{\Upsilon}^i)\big)\mathcal{Q}_{\nu_i}^{+}(\boldsymbol{\Upsilon}^1)\|_{\infty,\infty} \\
&\leq \max_{i\in[2,\dots,J]} \|\mathcal{Q}_{\nu_i}^{+}(\boldsymbol{\Upsilon}^i)\mathbf{1}_N\|_\infty^{J-1}\|\mathcal{Q}_{\nu_1}^{+}(\boldsymbol{\Upsilon}^1)\|_{\infty,\infty} \\
&= \kappa^{-1}\big(\kappa\frac{\nu}{\sqrt{2}}\big)^J,
\end{aligned}$$

where $\ell_{\infty,\infty}$-bound of the multiplication of matrices is recasted as mutliplication of the individual $\ell_{\infty,\infty}$ bound. □

It is important to note that both these Lemmas introduce bounds on the matrix that are extremely general. In that sense, because they consider a worst-case upper-bound that ignores the structure that makes up these factorized model, they are extremely loose. The next results will highlight their impact and show how, for specific models (*e.g.*, the FFT) how their developed bounds can be tightened.

The bound on the discrepancy between the high-resolution back-projection and the QPBP reconstruction, using the factorized model, is first studied in a non-uniform setting.

We first developed a non uniform bound for a given vector $\boldsymbol{x} \in \tilde{\Sigma}_s^N$,

we then extend this property to all s-sparse vector using a union bound argument and a covering.

Theorem 8.7. *Given ℓ_∞-bounded, ρ-repeated, and possibly quantized measurement $\tilde{\boldsymbol{S}}\tilde{\boldsymbol{z}} \in \mathbb{C}^{\rho N}$ such that $\|\boldsymbol{z}\|_\infty \leq \frac{\epsilon}{\sqrt{2}}$, a sub-sampling matrix $\tilde{\boldsymbol{S}} \in \mathbb{R}^{\rho N \times \rho N}$ as in (8.8) such that $\mathrm{Tr}\,(\tilde{\boldsymbol{S}}) = \rho\mu N = m$, given the quantization and the factorized model defined in (8.7), one can bound, for all support sets $\mathcal{T}$ such that $|\mathcal{T}| \leq 2s$, the following expression :*

$$\mathbb{P}\{\frac{1}{m}\|(\boldsymbol{P}^\rho(\prod_{i=1}^{J}\mathrm{diag}_\rho(\boldsymbol{\Upsilon}^i) - \prod_{i=1}^{J}\mathcal{Q}_\nu^+(\mathrm{diag}_\rho(\boldsymbol{\Upsilon}^i)))\tilde{\boldsymbol{S}}\tilde{\boldsymbol{z}})_\mathcal{T}\|_2 \geq \sqrt{2seJ}\gamma\epsilon B\} \leq N\exp\left(-2cm\gamma^2\right)$$

with $B = \left(\nu\sqrt{\frac{N\kappa}{2}}\right)^J$, and $\boldsymbol{P}^\rho := (\boldsymbol{I}_N, \ldots, \boldsymbol{I}_N) \in \mathbb{C}^{N\times\rho N}$.

This non-uniform bound highlights the benefits of using Theorem 8.4 that is able to leverage the factorized model compared to the classical Hoeffding inequality directly. Indeed, the degradation term is here shown to behave as $\mathcal{O}(m^{-\frac{1}{2}})$ in the non-uniform setting. This bound clearly outperforms the naive bound that used only the $\kappa\rho$ random variables coming from the last sub-matrices $\mathcal{Q}_{\nu_J}^+(\boldsymbol{\Upsilon}^J)$.

The term B in the upper-bound in Theorem 8.7 might seem to explode as the number of sub-matrices J increases. One must, however, note that the upper-bound on B developped in Lemma 8.5 is extremely pessimistic as it ignores any structure that $\boldsymbol{\Upsilon}^i$ might have. Indeed, for specific factorizable model such as the FFT, the bound on the spectral norm of the quantized version of $\boldsymbol{\Upsilon}^i$ can be dramatically lower. The spectral norm of the non-quantized sub-matrices for the FFT is $\|\boldsymbol{\Upsilon}^i\| = \sqrt{\kappa}$, which in turns gives $B \leq \sqrt{\kappa}^J$. Given that for the radix-2 algorithm [CT65] $\kappa = 2$ and $J = \log_2(N)$, B it then equals to $\sqrt{N}$. While the sub-matrices are quantized to 1-bit with an additive dither, they can be expected to partly retain the structure of the original matrices and thus not systematically reach the pessimistic bounds in Lemma 8.5 and Lemma 8.6.

The results established in Theorem 8.7 are only valid for one vector in

$\tilde{\Sigma}_s^N$, the rest of this section focuses on extending this non-uniform bound to all complex s-sparse vectors. For the sake of conciseness, we now omit the explicit factorized representation of the back-projection and of the over-sampling that samples ρ repetition of the measurements. The (possibly repeated and sub-sampled) measurements are denoted by $\tilde{\boldsymbol{z}} \in \mathbb{C}^m$ which are associated to a measurement matrix $\tilde{\boldsymbol{\Phi}} := [\boldsymbol{\Phi}^1, \dots, \boldsymbol{\Phi}^\rho] \in \mathbb{C}^{m \times N}$, with the super-script on the different $\boldsymbol{\Phi}$ representing the ρ different possible sub-sampling applied to each measurement matrix. The quantized and factorized back-projection is represented as $\tilde{\boldsymbol{\Psi}}^H \in \mathbb{C}^{N \times m}$, where the ρ repeated and factorized BP (with the appropriate sub-sampling) are quantized with indepedent dithers.

Theorem 8.8. *For all unit s-sparse vectors $x \in \tilde{\Sigma}_s^N$, given a measurement matrix $\tilde{\boldsymbol{\Phi}} \in \mathbb{C}^{m \times N}$, with $\|\tilde{\boldsymbol{\Phi}}\boldsymbol{x}\|_\infty \leq \frac{\epsilon}{2}$, whose back-projection can be factorized by J submatrices whose line are κ-sparse, where all components of these sub-matrices are bounded by $\frac{\nu}{2}$. Furthermore, the matrix $\tilde{\boldsymbol{\Phi}}$ follows the (ℓ_2, ℓ_2)-RIP; the 1-bit quantized version of this factorized back-projection, following the structure defined in (8.7), is denoted by $\tilde{\boldsymbol{\Psi}}^H$. For all support set $\mathcal{T}$ such that $|\mathcal{T}| \leq 2s$. It can be shown that, with a probability of failure exceeding $N \exp\left(-\frac{c}{2}\gamma^2 m\right)$,*

$$\tfrac{1}{m}\|((\tilde{\boldsymbol{\Phi}}^H - \tilde{\boldsymbol{\Psi}}^H)\tilde{\boldsymbol{\Phi}}\boldsymbol{x})_{\mathcal{T}}\| \leq \sqrt{8seJ}\epsilon B\gamma,$$

Provided $m \geq \frac{2s}{c}\gamma^{-2}\big(\log(\frac{eN}{s}) + 2\log(1+\frac{2}{\alpha})\big)$ and with $\alpha = \eta\frac{\kappa}{4\sqrt{s}(\frac{\kappa\nu}{\sqrt{2}})^J}$ being a constant dependent on the problem.

This last condition on m also gives the behaviour of the reconstruction algorithm, *i.e.*, $\gamma = \mathcal{O}(sm^{-\frac{1}{2}})$. We observe again an extra $\sqrt{s}$ factor. We finish this section with the bound when reconstructing from quantized and dithered measurements.

Theorem 8.9. *For all unit s-sparse vectors $x \in \tilde{\Sigma}_s^N$, given a measurement matrix $\tilde{\boldsymbol{\Phi}} \in \mathbb{C}^{m \times N}$, with $\|\tilde{\boldsymbol{\Phi}}\boldsymbol{x}\|_\infty \leq \frac{\epsilon}{2}$, whose back-projection can be fac-*

torized by J submatrices whose line are κ-sparse, where all components of these sub-matrices are bounded by $\frac{\nu}{2}$. Furthermore, the matrix $\tilde{\mathbf{\Phi}}$ follows the (ℓ_2, ℓ_2)-RIP; the 1-bit quantized version of this factorized back-projection, following the structure defined in (8.7), is denoted by $\tilde{\mathbf{\Psi}}^H$. For all support set $\mathcal{T}$ such that $|\mathcal{T}| \leq 2s$. With probability exceeding $2N \exp\left(-\frac{c}{2} m\gamma^2\right)$ with it can be shown that :

$$\frac{1}{m}\|((\tilde{\mathbf{\Phi}}^H - \tilde{\mathbf{\Psi}}^H)\mathcal{Q}_\epsilon^+(\tilde{\mathbf{\Phi}}\boldsymbol{x}))_\mathcal{T}\| \leq \sqrt{2s}\epsilon\gamma(\sqrt{eJ}B + \frac{8}{3\kappa}c(\frac{\nu\kappa}{\sqrt{2}})^J)$$

provided $m \geq \frac{2s}{c}\gamma^{-2}(\log(\frac{eN}{s}) + 2(1 + \frac{2}{\alpha}))$, with $\alpha = \sqrt{\frac{2}{9m}}c\gamma^2\epsilon$.

This last bound also provides the same behaviour for the discrepancy between the PBPQ and QPBPQ of $\mathcal{O}(\frac{s}{\sqrt{m}})$.

These theorems show that quantizing the individual elements of a factorized back-projection can still provides a bounded reconstruction error while maintaining their structure that allow for fast reconstructions.

8.6 Simulation and Discussion

We assess the quality of the developed schemes by performing Monte-Carlo simulations. We carried out 100 runs for different numbers of measurements and sparsity levels. The sparse vectors as well as the measurement matrix are generated similarly to Part II.

We use the following convention to represent the different quantization scheme : the classic PBP with linear measurements (◐); classic PBP with Quantized measurements (◑); Quantized PBP with linear measurements (◐); and finally Quantized PBP with Quantized measurements (◐).

In Fig. 8.2, we compare the proposed QPBPQ scheme in terms of ℓ_2-reconstruction error against PBP and PBPQ for two sparsity level, $s = 2$ and $s = 10$, with a measurement matrix corresponding to a randomly (sub/over)-sampled Fourier transform applied simply as a matrix multiplication. As indicated by the developed proofs in Theorems 8.2 and 8.3, the error of QPBPQ does scale as $\mathcal{O}(m^{-\frac{1}{2}})$ and only suffers from a con-

stant loss in dB compared to PBPQ. Interestingly, it seems that the loss of resolution in the measured signal has more impact on the performances than lowering the resolution of the back-projection, as shown by the shift between PBP, PBPQ and QPBPQ.

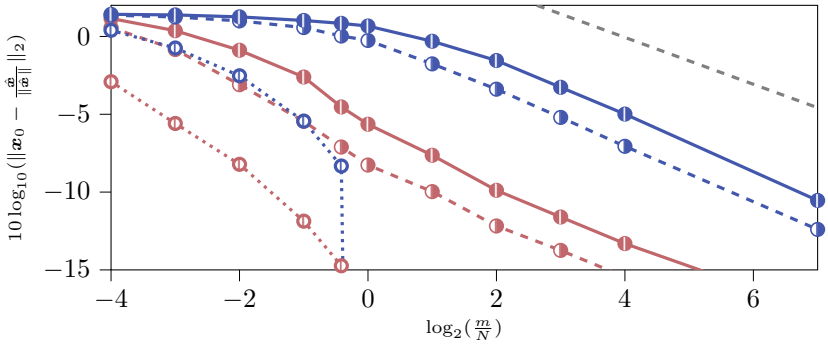


Figure 8.2: $\|\boldsymbol{x} - \frac{\hat{\boldsymbol{x}}}{\|\hat{\boldsymbol{x}}\|_2}\|_2$ in dB, for different numbers of measurements ($\log_2(\frac{m}{N})$), the dotted curves are the classic PBP; the dashed, PBPQ and the solid, QPBPQ. The colours represent the sparsity, $s = 2$ for **red** and $s = 10$ for **blue**. The dashed grey line represents the decrease rate of $\mathcal{O}(m^{-\frac{1}{2}})$.

In Fig. 8.3, we also see that using complex Gaussian matrices yields similar results, albeit requiring a larger number of measurements to reach the same performances compared to Fourier transforms. This, intuitively makes sense; the Fourier matrix is highly structured and each component has the same amplitude, the 1-bit quantization is able to partly retain these properties, whereas, complex Gaussian elements have no structure and vary greatly in amplitude.

We now examine the behaviour of the factorized back-projection studied in Section 8.5. In this simulations, we only focus on the Fourier transform and its factorized representation into $\log_2(N)$ sub-matrices [CT65] as it is one of most ubiquitous factorized back-projection used in modern signal processing.

The model in (8.7) decouples the number of measurements m into two terms, the sub-sampling coefficient μ and the number of repeated acquisitions ρ. In Fig. 8.4, the number of repetitions is varied from $\rho = 1$ to $\rho = 32$ and μ is set to 1 to remove the effect of the sub-sampling. One can observe that QPBP and QPBPQ have the same slope as the curve of PBPQ, which

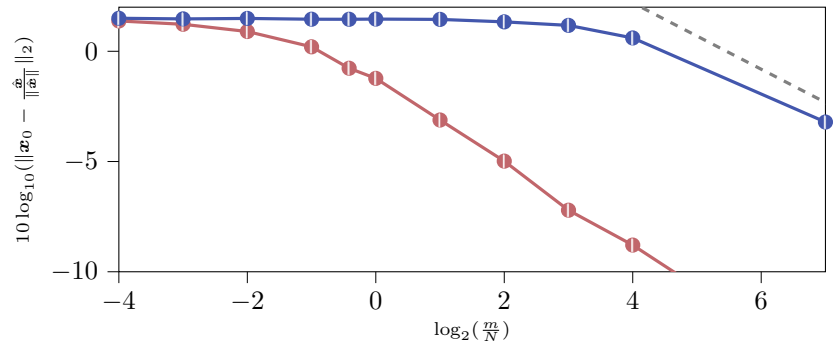


Figure 8.3: $\|\boldsymbol{x} - \frac{\hat{\boldsymbol{x}}}{\|\hat{\boldsymbol{x}}\|_2}\|_2$ in dB, for different number of measurements ($\log_2(\frac{m}{N})$), for different schemes with a sparsity of $s = 4$, namely QPBPQ with dithering in **red** for Fourier matrices and the QPBPQ with dithering for complex Gaussian matrices in **blue**.

follows $\mathcal{O}(m^{-\frac{1}{2}})$. As predicted by the theory in Theorems 8.8 and 8.9, the degradation imparted by the quantization of the factorized back-projection follows the behaviour of the reconstruction error of the high resolution processing, thus only shifting the curves in Fig. 8.4.

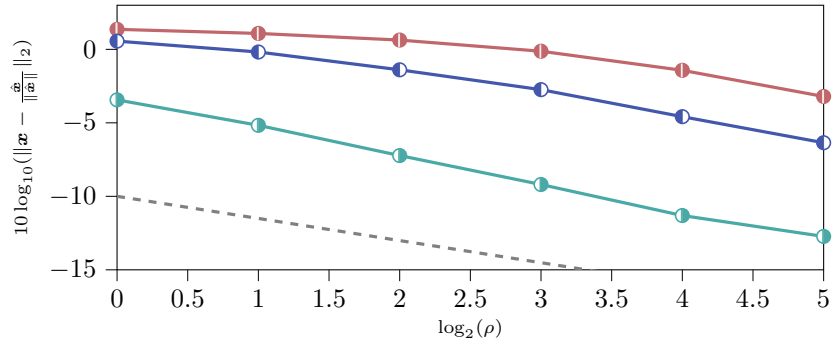


Figure 8.4: $\|\boldsymbol{x} - \frac{\hat{\boldsymbol{x}}}{\|\hat{\boldsymbol{x}}\|_2}\|_2$ in dB for the BP performed as a factorized model, for different values of repetition ρ ($\log_2(\rho)$), for different schemes with a sparsity of $s = 4$ and $\mu = 1$, namely QPBPQ in **red**; PBP in **yellow**; QPBP in **blue**; PBPQ in **green**; the dashed **gray** line represents $\mathcal{O}(\rho^{-\frac{1}{2}})$.

Let us now study the behaviour of the quantized scheme when the measurements are sub-sampled, *i.e.*, for $\mu \leq 1$. In Fig. 8.5, one can observe that although each element of $\hat{\boldsymbol{x}}$ is a sum of at most κ random variables coming from the last sub-matrix $\mathcal{Q}^+_{\nu_J}(\boldsymbol{\Psi}^J)$, ℓ_2-reconstruction of QPBPQ and QPBP

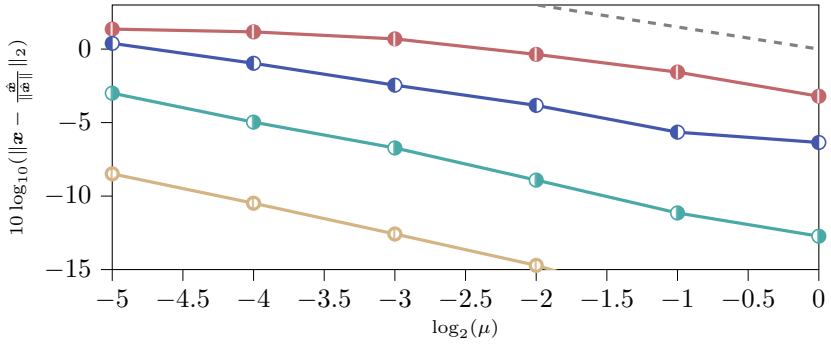


Figure 8.5: $\|\boldsymbol{x} - \frac{\hat{\boldsymbol{x}}}{\|\hat{\boldsymbol{x}}\|_2}\|_2$ in dB for the BP performed as a factorized model, for different values of sub-sampling μ ($\log_2(\mu)$), for different schemes with a sparsity of $s = 4$ and $\rho = 32$, namely QPBPQ in **red**; PBP in **yellow**; QPBP in **blue**; PBPQ in **green**; the dashed **gray** line represents $\mathcal{O}(\mu^{-\frac{1}{2}})$.

behaves as $\mathcal{O}(\mu^{-\frac{1}{2}})$ instead of $\mathcal{O}(\kappa^{-\frac{1}{2}})$. This highlights the tightness of the estimate provided by Remarks 5.5 & 5.7 (Uniform Bounds on Factors) from [Hua+20] and consequently the non-uniform results in Lemma 8.11.

The results presented in Fig. 8.4 and Fig. 8.5 show clearly that the factorized scheme indeed behaves as $\mathcal{O}(m^{-\frac{1}{2}})$, with $m = \rho\mu N$. This means that these factorized and quantized schemes are able to be computed efficiently in $\mathcal{O}(J\kappa)$ computations while maintaining a reconstruction bound that scales as $\mathcal{O}(m^{-\frac{1}{2}})$.

In Fig. 8.6 and Fig. 8.7, we compare different schemes with the same sparsity level of $s = 4$, with and without a random dither added before the quantization of the elements of the BP. Similarly to what was shown in [XJ19; Feu+18a; Feu+18b], we see the dither plays in capital role in the obtained performances. Indeed, while the dithered scheme continues to scale down when m increases, the scheme with the deterministic back-projection seems to slowly saturates. In these figures, the coefficient μ is varied when $\log_2(\frac{m}{N}) \leq 0$ with $\rho = 1$, beyond this point, μ is fixed to one and the number of repetitions ρ is varied. In Fig. 8.6, using a deterministic and quantized back-projection operator, is tantamount in the context of Fourier transform to reconstructing Fourier based measurements with complex Walsh matrices that are their 1-bit equivalent, *i.e.*, $\boldsymbol{W} = \mathcal{Q}_\nu(\boldsymbol{F})$. In this context, it is clear for the case of matrix multiplication in Fig. 8.6 that the dithering is

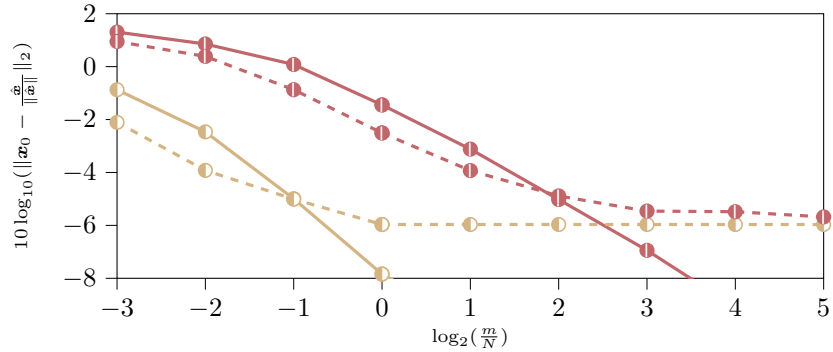


Figure 8.6: Comparison using $\|\boldsymbol{x} - \frac{\hat{\boldsymbol{x}}}{\|\hat{\boldsymbol{x}}\|_2}\|_2$ in dB between the Quantized BP performed as a matrix multiplication with additive dithering (solid) and without (dashed), for different number of measurements ($\log_2(\frac{m}{N})$), with a sparsity of $s = 4$, namely QPBPQ in **red**; QPBP in **yellow**.

necessary to achieve the best performances.

For the Factorized BP in Fig. 8.7, we also observe that the non dithered schemes saturates. In fact, they saturates at a value that is extremely high ($\approx 1dB$), to the point where one could say that the reconstruction fails. Theorems 8.8 and 8.9 rely heavily on the added dither as it provides $\mathbb{E}\{\prod_i^J \mathcal{Q}_\nu^+(\boldsymbol{\Upsilon}^i)\} = \prod_i^J \boldsymbol{\Upsilon}^i$. Without the dither, this does not hold and the resulting $\tilde{\boldsymbol{\Psi}}^H$ cannot be linked to its high resolution counter-part.

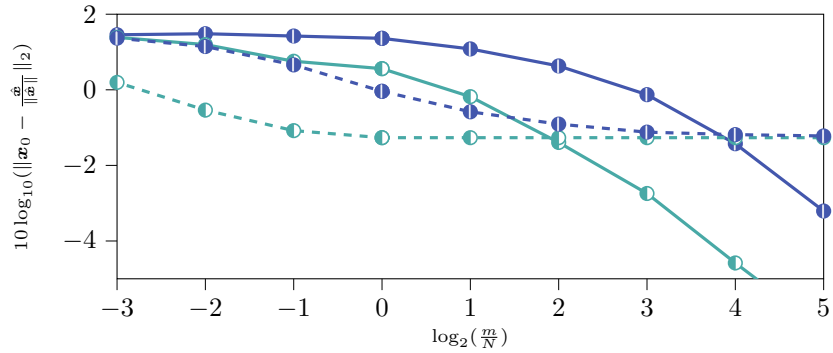


Figure 8.7: Comparison using $\|\boldsymbol{x} - \frac{\hat{\boldsymbol{x}}}{\|\hat{\boldsymbol{x}}\|_2}\|_2$ in dB between the Quantized BP performed as a factorized model with additive dithering (solid) and without (dashed), for different number of measurements ($\log_2(\frac{m}{N})$), with a sparsity of $s = 4$, namely QPBPQ in **blue**; QPBP in **green**.

We finish this section by comparing the two studied quantized reconstruction processes, namely the application of the back-projection as a matrix-vector multiplication and the use of an efficient factorized representation. Instead of comparing their ℓ_2-reconstruction with respect to the number of measurement m, we chose to compare them using an estimate of the complexity in terms of number of operation required by the estimation process. On the one hand, the number of computations required to multiply a $N \times m$ matrix to a m vectors is $\mathcal{O}(mN)$, on the other hand, in the case of the FFT, the complexity is $\mathcal{O}(m \log_2(N))$. In Fig. 8.8, the different schemes are compared. We see that the two quantization procedures for the BP give similar reconstruction for the same number of operations. Given the specificity of implementing algorithms and arithmetical functions on a processing unit like an FGPA, developing a more refined metric that would highlight the difference between these scheme is out of the scope of this work. The complexity still provides a glimpse of the gain that the factorized model could provide with a dedicated hardware implementation.

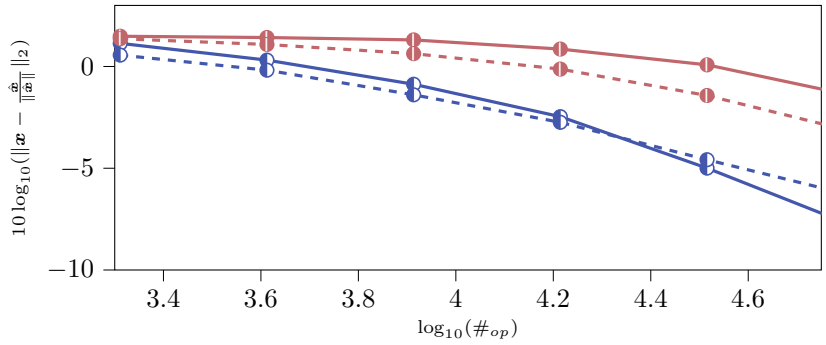


Figure 8.8: Comparison using $\|\boldsymbol{x} - \frac{\hat{\boldsymbol{x}}}{\|\hat{\boldsymbol{x}}\|_2}\|_2$ in dB between the BP performed as a matrix multiplication (solid) and factorized model (dashed), for different computational complexity $\#_{op}$ ($\log_2(\#_{op})$), for different schemes with a sparsity of $s = 4$, namely QPBPQ in **red**; PBP in **yellow**; QPBP in **blue**.

8.7 Discussion

In this chapter, we showed that quantizing the matrix used for the back-projection in the PBP algorithm can still provide interesting reconstruction

guarantees. We showed, both theoretically and through Monte-Carlo simulations, that directly quantizing the back-projection operator to 1-bit with an additive dither provides only a constant degradation that can be compensated by slightly increasing the number of measurements. These results were then extended to the special case of back-projection operators that have a fast matrix-vector multiplication thanks to a factorized representation. We developed uniform bounds that apply to all matrices that have both a factorized representation and follow the RIP property. We showed that, regardless of the acquisition procedure (*e.g.*, linear or quantized), the reconstruction obtained by this extremely coarse processing behave similarly to its high resolution counterpart (*i.e.*, $\mathcal{O}(m^{-\frac{1}{2}})$) up to a constant, but required more measurements than the simple matrix quantization.

Future works regarding this area of compressive sensing are numerous. The bound could be tightened in order to change the dependency on the sparsity from $\mathcal{O}(s)$ to $\mathcal{O}(\sqrt{s})$. The results in the second part for the factorizable BP are introduced in a general setting. The results could thus be focused on specific applications and matrices, and yield tighter bounds. The developed theory as well as the simulations only considered a noiseless case, while the extension to noisy measurements before the 1-bit quantization with additive dithering is still an open question in the classic PBP, result for linear measurements could be easily obtained. One could also extend this scheme beyond the PBP algorithm. Indeed, the factorized model studied in Sec.8.5 can also be seen as a connected network of nodes where the weights are quantized to one bit. So the extension to neural networks and deep unfolding could be of interest [MLE21; Mer+16]. Finally, the objective of this chapter was to study reconstruction schemes that could be efficiently implemented in hardware. Now that the first theoretical guarantees have been established, a more in-depth study on the actual implementation and potential gain of the simplified 1-bit BP must be performed.

8.8 Proofs

This contains the different proofs of the theorems and lemmas that were developed in this chapter.

For any bounded s-sparse vector $\boldsymbol{x}$ whose closest point in $\mathcal{J}_\rho$ is $\boldsymbol{u}$, one can

leverage Lemma 6.1 in [XJ19] to bound the number of non-zero components in the vector $\mathcal{Q}_\epsilon^+(\boldsymbol{\Phi x}) - \mathcal{Q}_\epsilon^+(\boldsymbol{\Phi u})$ (where they are both dithered with the same random dither) to a small fraction of m. Because this lemma is used multiple times in the different proofs, we state this Lemma and adapt it to the considered complex 1-bit dithered measurements.

Lemma 8.10 (Lemma 6.1 [XJ19])**.** *Given $\boldsymbol{a} \in \mathbb{R}^m$, $\epsilon > 0$, $0 < \rho < \frac{\delta}{2}$, $\boldsymbol{\xi} \sim \mathcal{U}^m([0,\delta])$, we denote by $\underline{\mathsf{A}_i}(\cdot) := \mathcal{Q}(\cdot + \boldsymbol{\xi})$ ($i \in [m]$), and define the discrete random variable (associated with the randomness of $\boldsymbol{\xi}$)*

$$Z = Z(\boldsymbol{a} + \rho\mathbb{B}_{\ell_\infty}^m) := |\{i : \underline{\mathsf{A}_i}(\cdot) \notin \mathsf{C}^0(\boldsymbol{a}) + \rho\mathbb{B}_{\ell_\infty}^m)\}| \in \{0, \dots, m\}$$

i.e., *Z counts the components of $\underline{\mathsf{A}}$ that are discontinuous over $\boldsymbol{a}+\rho\mathbb{B}_{\ell_\infty}^m$* (i.e., *having at least two distinct values over this set). The random variable Z has a binomial distribution with m trials and a probability of success $p := \frac{2\rho}{\delta}$*, i.e., *$Z sim Bin(m,p)$. Therefore, $\mathbb{E}Z = mp = m\frac{2\rho}{\delta}$, $\frac{1}{m}(\mathbb{E}Z^2 - (\mathbb{E}Z)^2 = p(1-p) =: \sigma^2 < p$ and*

$$\mathbb{P}[Z \geq m\frac{2\rho}{\delta} + \epsilon] \leq \exp\left(-\frac{1}{2}\frac{3m\epsilon^2}{3\sigma^2 + \epsilon}\right)$$

In particular, setting $\epsilon = p > \sigma^2$ provides

$$\mathbb{P}[Z \geq m\frac{4\rho}{\delta}] \leq \exp\left(-m\frac{3\rho}{4\delta}\right).$$

In the following remark, we adapt Lemma 8.10 from its general setting of any vector $\boldsymbol{a} \in \mathbb{R}^m$ to the specific setting of $\boldsymbol{\Phi x} \in \mathbb{C}^m$ complex measurements.

Remark 8.1 (Adaptation of Lemma 6.1 [XJ19])**.** Considering the complex quantization, one can count the number of differing bits in the complex domain of two different measurements vectors in $\mathbb{C}^m$ by recasting in the real domain, *i.e.*, $\mathbb{R}^{2m}$. Furthermore, the ρ-covering on defined in Lemma 6.4

and the RIP property, $\boldsymbol{\Phi}$ which upper-bounds $\|\boldsymbol{\Phi x}\|_\infty \leq \|\boldsymbol{\Phi x}\|_2 \leq \sqrt{2m}\rho$, allows to restate Lemma 8.10 as

$$\|\mathcal{Q}_\epsilon^+(\boldsymbol{\Phi u}) - \mathcal{Q}_\epsilon^+(\mathcal{Q}\boldsymbol{\Phi}(\boldsymbol{u}+\boldsymbol{r}))\|_0 \leq 8\sqrt{2}m^{\frac{3}{2}}\frac{\rho}{\epsilon},$$

with a probability of failure upper-bounded by $\exp(-\frac{3}{\sqrt{2}}m^{\frac{3}{2}}\frac{\rho}{\epsilon})$.

Theorem 8.2. *For all s-sparse vector $\boldsymbol{x}$, with $\|\boldsymbol{x}\|_2 \leq 1$, a measurement matrix $\boldsymbol{\Phi} \in \mathbb{C}^{m\times N}$ that follows the (ℓ_2,ℓ_2)-RIP$(\delta, 2s)$, considering furthermore that $\|\boldsymbol{\Phi x}\|_\infty \leq \frac{\epsilon}{2}$ and that $\|\boldsymbol{\Phi}\|_{\infty,\infty} \leq \frac{\nu}{2}$, for all support sets $\mathcal{T}$ of size $2s$ and using a modified back-projection operator defined as $\boldsymbol{\Psi}^H := \mathcal{Q}_\nu^+(\boldsymbol{\Phi}^H)$, one can upper-bound, with a probability exceeding $1 - 2N\exp\left(\frac{-m\gamma^2}{32s\epsilon^2\nu^2}\right)$,*

$$\|((\boldsymbol{\Psi}-\boldsymbol{\Phi})^H\boldsymbol{\Phi x})_\mathcal{T}\|_2 \leq 3m\gamma,$$

provided $m \geq 16s^2\epsilon^2\nu^2\gamma^{-2}\big(\log(\frac{eN}{s}) + 2\log(1+\frac{2\sqrt{2s}\nu}{\gamma})\big)$.

Proof. Lemma 8.1 is only valid for one fixed vector $\boldsymbol{x} \in \tilde{\Sigma}_s^N$. We use a ρ-covering $\mathcal{J}_\rho$ to extend it to all complex s-sparse vectors in $\tilde{\Sigma}_s^N$. Similarly to Theorem 6.7 in Chapter 6, by union bound, Lemma 8.1 holds for all $\boldsymbol{u} \in \mathcal{J}_\rho$ with probability exceeding $1 - 2N\exp(\log|\mathcal{J}_\rho| - \frac{m\gamma^2}{16s\epsilon^2\nu^2})$. We then extend this inequality from all vectors in the covering (*i.e.*, $\boldsymbol{u} \in \mathcal{J}_\rho$) to all s-sparse vectors $\boldsymbol{x} \in \tilde{\Sigma}_S^N$ by leveraging the RIP and the properties of the covering found in Remark 6.1 and Lemma 6.4.

$$\begin{aligned}\|\big((\boldsymbol{\Psi}-\boldsymbol{\Phi})^H\boldsymbol{\Phi x}\big)_\mathcal{T}\|_2 &\leq \|\big((\boldsymbol{\Psi}-\boldsymbol{\Phi})^H\boldsymbol{\Phi u}\big)_\mathcal{T}\|_2 + \|\big((\boldsymbol{\Psi}-\boldsymbol{\Phi})^H\boldsymbol{\Phi r}\big)_\mathcal{T}\|_2 \\ &\leq m\gamma + \|\big((\boldsymbol{\Psi}-\boldsymbol{\Phi})^H\boldsymbol{\Phi r}\big)_\mathcal{T}\|_2,\end{aligned}$$

with a probability of failure upper-bounded by $|\mathcal{J}_\rho|2N\exp(-\frac{m\gamma^2}{16s\epsilon^2\nu^2})$. One can bound the second term, using the fact that $\forall i \in [N]\ |\boldsymbol{\Phi}_{ij}^H - \boldsymbol{\Psi}_{ij}^H| \leq \sqrt{2}\nu$,

which gives us,

$$\begin{aligned}
&\leq m\gamma + \sqrt{|\mathcal{T}|}\|(\boldsymbol{\Psi}^H - \boldsymbol{\Phi}^H)\boldsymbol{\Phi}\boldsymbol{r}\|_\infty \\
&\leq m\gamma + \sqrt{|\mathcal{T}|}\sqrt{2}\nu\|\boldsymbol{\Phi}\boldsymbol{r}\|_1 \\
&\leq m\gamma + 2\sqrt{ms}\nu\|\boldsymbol{\Phi}\boldsymbol{r}\|_2
\end{aligned}$$

Using the property of the ρ-covering and the RIP of $\boldsymbol{\Phi}$ one can finally bound,

$$\|\big((\boldsymbol{\Psi} - \boldsymbol{\Phi})^H\boldsymbol{\Phi}\boldsymbol{x}\big)_{\mathcal{T}}\|_2 \leq m\gamma + 2\sqrt{2s}\nu m\rho,$$

with probability of failure upper-bounded by $2N\exp(\log|\mathcal{J}_\rho| - \frac{m\gamma^2}{16s\epsilon^2\nu^2})$.

Setting $\rho = \frac{\gamma}{\nu\sqrt{2s}}$, the inequality becomes

$$\|((\boldsymbol{\Psi} - \boldsymbol{\Phi})^H\boldsymbol{\Phi}\boldsymbol{x})_{\mathcal{T}}\|_2 \leq 3m\gamma.$$

Finally, the probability of failure can be upper-bounded by

$$2N\exp(\log|\mathcal{J}_{\frac{\gamma}{\nu\sqrt{2s}}}| - \frac{m\gamma^2}{16s\epsilon^2\nu^2}) \leq 2N\exp(-\frac{m\gamma^2}{32s\epsilon^2\nu^2})$$

Provided, using the bounded size of $|\mathcal{J}_\rho|$ in Lemma 6.4, that

$$16s^2\epsilon^2\nu^2\gamma^{-2}\big(\log(\frac{eN}{s}) + 2\log(1 + \frac{2\sqrt{2s}}{\gamma})\big) \leq m.$$

□

Theorem 8.3. *For all s-sparse vector $\boldsymbol{x} \in \tilde{\Sigma}_s^N$, with $\|\boldsymbol{x}\|_2 \leq 1$, a measurement matrix $\boldsymbol{\Phi} \in \mathbb{C}^{m\times N}$ that follows the (ℓ_2, ℓ_2)-RIP$(\delta, 2s)$, considering furthermore that $\|\boldsymbol{\Phi}\boldsymbol{x}\|_\infty \leq \frac{\epsilon}{2}$ and that $\|\boldsymbol{\Phi}\|_{\infty,\infty} \leq \frac{\nu}{2}$, for all support set $\mathcal{T}$ of size $2s$ and using a modified back-projection operator defined as $\boldsymbol{\Psi}^H := \mathcal{Q}_\nu^+(\boldsymbol{\Phi}^H)$, one can upper-bound with a probability exceeding $1 - 2N\exp\big(\frac{-m\gamma^2}{32s\epsilon^2\nu^2}\big)$,*

$$\|((\boldsymbol{\Psi} - \boldsymbol{\Phi})^H\mathcal{Q}_\lambda^+(\boldsymbol{\Phi}\boldsymbol{x}))_{\mathcal{T}}\|_2 \leq 3m\gamma,$$

provided $m \geq 32\epsilon^2\nu^2\gamma^{-2}s^2\big(\log(\frac{eN}{s}) + 2\log(1 + 24\frac{\epsilon\nu^2\sqrt{2m}}{\gamma^2})\big)$.

Proof. We follow the same steps as in Theorem 8.2, by defining a ρ-covering $\mathcal{J}_\rho$ such that $\boldsymbol{x} = \boldsymbol{u} + \boldsymbol{r}$, with $\boldsymbol{x} \in \tilde{\Sigma}_s^N$, $\boldsymbol{u} \in \mathcal{J}_\rho$ and $\|\boldsymbol{r}\|_2 \leq \rho$, and extending the non-uniform result by a union bound on $\boldsymbol{u}$. One can bound the discrepancy between QPBPQ and PBPQ as follows

$$\begin{aligned}\|((\boldsymbol{\Psi}-\boldsymbol{\Phi})^H \mathcal{Q}_\epsilon^+(\boldsymbol{\Phi}\boldsymbol{x}))_\mathcal{T}\|_2 \leq & \|((\boldsymbol{\Psi}-\boldsymbol{\Phi})^H \mathcal{Q}_\epsilon^+(\boldsymbol{\Phi}\boldsymbol{u}))_\mathcal{T}\|_2 \\ & + \|\big((\boldsymbol{\Psi}-\boldsymbol{\Phi})^H(\mathcal{Q}_\epsilon^+(\boldsymbol{\Phi}\boldsymbol{x}) - \mathcal{Q}_\epsilon^+(\boldsymbol{\Phi}\boldsymbol{u}))\big)_\mathcal{T}\|_2 \\ \leq & m\gamma + \|\big((\boldsymbol{\Psi}-\boldsymbol{\Phi})^H(\mathcal{Q}_\epsilon^+(\boldsymbol{\Phi}\boldsymbol{x}) - \mathcal{Q}_\epsilon^+(\boldsymbol{\Phi}\boldsymbol{u}))\big)_\mathcal{T}\|_2 \end{aligned} \tag{8.10}$$

with a probability of failure upper-bounded by $|\mathcal{J}_\rho| 2N \exp\big(\frac{-m\gamma^2}{16 s\epsilon^2\nu^2}\big)$.

Focusing on the second term of (8.10), one can subsequently bound the following

$$\begin{aligned}\|((\boldsymbol{\Psi}-\boldsymbol{\Phi})^H(\mathcal{Q}_\epsilon^+(\boldsymbol{\Phi}\boldsymbol{x}) - \mathcal{Q}_\epsilon^+(\boldsymbol{\Phi}\boldsymbol{u})))_\mathcal{T}\|_2 & \\ \leq \sqrt{|\mathcal{T}|}\max_i |(\boldsymbol{\Psi}-\boldsymbol{\Phi})_i^H(\mathcal{Q}_\epsilon^+(\boldsymbol{\Phi}\boldsymbol{x}) - \mathcal{Q}_\epsilon(\boldsymbol{\Phi}\boldsymbol{u}))| & \\ \leq \sqrt{|\mathcal{T}|}\sqrt{2}\nu\|\mathcal{Q}_\epsilon^+(\boldsymbol{\Phi}\boldsymbol{x}) - \mathcal{Q}_\epsilon^+(\boldsymbol{\Phi}\boldsymbol{u}))\|_1 & \\ \leq \sqrt{|\mathcal{T}|}2\nu\epsilon|\mathcal{D}|, & \end{aligned} \tag{8.11}$$

where $\mathcal{D}$ is the support set of index where the two quantized signals lie in different quadrants. Using Remark 8.1, thanks to the results in [XJ19], for an element of the covering $\mathcal{J}_\rho$, we finally obtain

$$\mathbb{P}[|\mathcal{D}| \geq 2m\sqrt{2m}\frac{\rho}{\epsilon}] \leq \exp\big(-2m\frac{\sqrt{2m}3\rho}{4\epsilon}\big).$$

Combining this bound with (8.11) by extending the previous inequality to all element of the covering $\mathcal{J}_\rho$, the discrepancy between the reconstruction process is upper-bounded by

$$\|((\boldsymbol{\Psi}-\boldsymbol{\Phi})^H \mathcal{Q}_\epsilon^+(\boldsymbol{\Phi}\boldsymbol{x}))_\mathcal{T}\|_2 \leq m\gamma + 2\nu\epsilon(2m)^{\frac{3}{2}}\frac{\rho}{\epsilon}$$

with a probability of failure $\mathbb{P} \leq |\mathcal{J}_\rho|\big(2N\exp\big(\frac{-m\gamma^2}{16 s\epsilon^2\nu^2}\big) + \exp\big(-(2m)^{\frac{3}{2}}\frac{3\rho}{4\epsilon}\big)\big)$.

By setting, $\rho = \frac{\gamma^2}{24s\epsilon\nu^2\sqrt{2m}}$, we obtain

$$\|((\boldsymbol{\Psi}-\boldsymbol{\Phi})^H \mathcal{Q}_\epsilon^+(\boldsymbol{\Phi}\boldsymbol{x}))_\mathcal{T}\|_2 \leq m\gamma(1+\frac{1}{3\epsilon\nu}), \tag{8.12}$$

with a probability of failure $\mathbb{P} \leq |\mathcal{J}_\rho|(2N+1)\exp\left(\frac{-m\gamma^2}{8\epsilon^2\nu^2}\right)$. Where (8.12) is obtained thanks to the fact that $\gamma^2 \leq \gamma \leq 1$.

Finally, if

$$m \geq 32\epsilon^2\nu^2 s^2\gamma^{-2}\big(\log(\frac{eN}{s}) + 2\log(1+24\frac{\epsilon\nu^2\sqrt{2m}}{\gamma^2})\big),$$

then the probability of failure can be upper-bounded by $(2N+1)\exp\left(\frac{-m\gamma^2}{32\epsilon^2\nu^2}\right)$. □

Lemma 8.11. *Given possibly subsampled measurements $\boldsymbol{S}\boldsymbol{z}$, with $\boldsymbol{z} \in \mathbb{C}^N$, $\|\boldsymbol{z}\|_\infty \leq \frac{\epsilon}{\sqrt{2}}$, $\boldsymbol{S} \in \mathbb{N}^{N\times N}$ such that $\mathrm{tr}(\boldsymbol{S}) = \mu N = m$, the ℓ_2-degradation between the PBP and QPBP for a factorizable back-projection model $\boldsymbol{\Phi}^H := \prod_i^J \boldsymbol{\Upsilon}^i$ with $\max_i \|\boldsymbol{\Upsilon}^i\|_{\infty,\infty} \leq \frac{\nu}{2}$ where each $\boldsymbol{\Upsilon}^i \in \mathbb{C}^{N\times N}$ is κ-line sparse,* i.e., *$\forall i \in [J], \|\boldsymbol{\Upsilon}^i_j\|_0 \leq \kappa$, one can bound*

$$\mathbb{P}\{\|(\prod_{i=1}^J \boldsymbol{\Upsilon}^i - \prod_{i=1}^J \mathcal{Q}_\nu^+(\boldsymbol{\Upsilon}^i))\boldsymbol{S}\boldsymbol{z}\|_2 \geq \frac{\epsilon}{\sqrt{2}}\sqrt{\mu N}\left(\nu\sqrt{\frac{N\kappa}{2}}\right)^J t\} \leq N\exp\big(-\frac{t^2}{eJ}\big).$$

Proof. We start by upper-bounding the ℓ_2 norm of the matrix vector product by the spectral norm of the difference between the back-projection and the ℓ_2 norm of $\boldsymbol{z}$, then using the ℓ_∞ bound of $\boldsymbol{z}$, we can show that

$$\begin{aligned}\|(\prod_{i=1}^J \boldsymbol{\Upsilon}^i - \prod_{i=1}^J \mathcal{Q}_\nu^+(\boldsymbol{\Upsilon}^i))\boldsymbol{S}\boldsymbol{z}\|_2 &\leq \|\prod_{i=1}^J \boldsymbol{\Upsilon}^i - \prod_{i=1}^J \mathcal{Q}_\nu^+(\boldsymbol{\Upsilon}^i)\|\|\boldsymbol{S}\boldsymbol{z}\|_2 \\ &\leq \|\prod_{i=1}^J \boldsymbol{\Upsilon}^i - \prod_{i=1}^J \mathcal{Q}_\nu^+(\boldsymbol{\Upsilon}^i)\|\frac{\epsilon}{\sqrt{2}}\sqrt{\mu N}. \end{aligned} \tag{8.13}$$

The $\frac{1}{\sqrt{2}}$ factor in (8.13) is used for the ℓ_∞-bound on $\boldsymbol{z}$ to put the emphasis on the fact that the result applies for both the linear and one-bit quantized measurements. Indeed, if $\|\boldsymbol{\Phi x}\|_\infty \leq \frac{\epsilon}{2}$, then $\|\boldsymbol{\Phi x}\|_\infty \leq \|\mathcal{Q}_\epsilon^+(\boldsymbol{\Phi x})\|_\infty \leq \frac{\epsilon}{\sqrt{2}}$.

Using Lemma 8.5, the quantity B and v^2 required by Theorem 8.4 can be upperbounded by

$$B \leq \Big(\nu\sqrt{\frac{N\kappa}{2}}\Big)^J, \qquad\qquad v^2 \leq \frac{J}{2}.$$

These bounds allow us to directly bound the sprectral norm in (8.13) by

$$\|(\prod_{i=1}^{J}\boldsymbol{\Upsilon}^i - \prod_{i=1}^{J}\mathcal{Q}_\nu^+(\boldsymbol{\Upsilon}^i))\boldsymbol{S z}\|_2 \leq t\Big(\nu\sqrt{\frac{N\kappa}{2}}\Big)^J \frac{\epsilon}{\sqrt{2}}\sqrt{\mu N},$$

with a probability of failure $p_t \leq N\exp\big(-\frac{t^2}{eJ}\big)$. □

Theorem 8.7. *Given ℓ_∞-bounded, ρ-repeated, and possibly quantized measurement $\tilde{\boldsymbol{S}}\boldsymbol{z} \in \mathbb{C}^{\rho N}$ such that $\|\boldsymbol{z}\|_\infty \leq \frac{\epsilon}{\sqrt{2}}$, a sub-sampling matrix $\tilde{\boldsymbol{S}} := (\boldsymbol{S}^1, \ldots, \boldsymbol{S}^\rho)^T \in \mathbb{R}^{\rho N \times N}$ such that $\mathrm{Tr}\,(\tilde{\boldsymbol{S}}) = \rho\mu N = m$ which contains all the sub-sampling matrices $\boldsymbol{S}^h$, given the quantization and the factorized model defined in (8.7), one can bound, for all support sets $\mathcal{T}$ such that $|\mathcal{T}| \leq 2s$, the following expression :*

$$\mathbb{P}\{\frac{1}{m}\|(\boldsymbol{P}^\rho(\prod_{i=1}^{J}\mathrm{diag}_\rho(\boldsymbol{\Upsilon}^i) - \prod_{i=1}^{J}\mathcal{Q}_\nu^+(\mathrm{diag}_\rho(\boldsymbol{\Upsilon}^i)))\tilde{\boldsymbol{S}}\boldsymbol{z})_\mathcal{T}\|_2 \geq \sqrt{2seJ}\gamma\epsilon B\}$$
$$\leq N\exp\left(-2cm\gamma^2\right)$$

with $B = \big(\nu\sqrt{\frac{N\kappa}{2}}\big)^J$, and the matrix $\boldsymbol{P}_\rho := (\boldsymbol{I}_N, \ldots, \boldsymbol{I}_N) \in \mathbb{C}^{N\times\rho N}$.

Proof. Starting from the ℓ_2-degradation, we can write, thanks to the defi-

nition of $\boldsymbol{P}^{\rho}$, the following

$$\begin{aligned}\frac{1}{m}\|(\boldsymbol{P}^{\rho}(\prod_{i=1}^{J}\operatorname{diag}_{\rho}(\boldsymbol{\Upsilon}^{i})-\prod_{i=1}^{J}\mathcal{Q}_{\nu_i}^{+}(\operatorname{diag}_{\rho}(\boldsymbol{\Upsilon}^{i})))\tilde{\boldsymbol{S}}\boldsymbol{z})_{\mathcal{T}}\|_2\\ =&\frac{1}{m}(\sum_{l\in\mathcal{T}}|\sum_{h}^{\rho}(\prod_{i=1}^{J}\boldsymbol{\Upsilon}^{i}-\prod_{i=1}^{J}\mathcal{Q}_{\nu_{ih}}^{+}(\boldsymbol{\Upsilon}^{i}))_l\boldsymbol{S}^{h}\boldsymbol{z}|^2)^{\frac{1}{2}}\\ \leq&\frac{\sqrt{|\mathcal{T}|}}{m}\max_{l}|\sum_{h}^{\rho}(\prod_{i=1}^{J}\boldsymbol{\Upsilon}^{i}-\prod_{i=1}^{J}\mathcal{Q}_{\nu_{ih}}^{+}(\boldsymbol{\Upsilon}^{i}))_l\boldsymbol{S}^{h}\boldsymbol{z}|\end{aligned}\tag{8.14}$$

To upperbound the ℓ_2-degradation, one needs to find a bound on the sum of the ρ back-projections in (8.14). We first notice that $\forall h\in[\rho]$ and $\forall l\in[N]$, $(\prod_{i=1}^{J}\boldsymbol{\Upsilon}^{i}-\prod_{i=1}^{J}\mathcal{Q}_{\nu_{ih}}^{+}(\boldsymbol{\Upsilon}^{i}))_l\boldsymbol{S}^{h}\boldsymbol{z}$ is zero-mean and that

$$|\sum_{h}^{\rho}(\prod_{i=1}^{J}\boldsymbol{\Upsilon}^{i}-\prod_{i=1}^{J}\mathcal{Q}_{\nu_{ih}}^{+}(\boldsymbol{\Upsilon}^{i}))_l\boldsymbol{S}^{h}\boldsymbol{z}|\leq\|(\prod_{i=1}^{J}\boldsymbol{\Upsilon}^{i}-\prod_{i=1}^{J}\mathcal{Q}_{\nu_{ih}}^{+}(\boldsymbol{\Upsilon}^{i}))_l\boldsymbol{S}^{h}\boldsymbol{z}\|_2\tag{8.15}$$

Furthermore, Lemma 8.11 shows that the second term in (8.15) is sub-gaussian. Consequently, the ρ terms in (8.14) are also sub-gaussians and mean-zero. Upperbounding a sum of ρ sub-gaussian random variables can be achieved using the *General Hoeffding's inequality* in [Ver18, Theorem 2.6.2]. The degradation is then upper-bounded by :

$$\frac{1}{m}\|(\boldsymbol{P}^{\rho}(\prod_{i=1}^{J}\operatorname{diag}_{\rho}(\boldsymbol{\Upsilon}^{i})-\prod_{i=1}^{J}\mathcal{Q}_{\nu_i}^{+}(\operatorname{diag}_{\rho}(\boldsymbol{\Upsilon}^{i})))\tilde{\boldsymbol{S}}\boldsymbol{z})_{\mathcal{T}}\|_2\leq\frac{\sqrt{|\mathcal{T}|}}{m}t,$$

with a probability of failure $p_t\leq N\exp\Big(-\frac{2ct^2}{\mu\rho NeJ\epsilon^2\left(\nu\sqrt{\frac{N\kappa}{2}}\right)^{2J}}\Big)$, with c being an absolute constant $c\in\mathbb{R}_+$.

A change of variable $\gamma=\frac{t}{m\epsilon\left(\nu\sqrt{\frac{N\kappa}{2}}\right)^{J}\sqrt{eJ}}$, finally gives the desired results. The ℓ_2-discrepancy is bounded by

$$\begin{aligned}\frac{1}{m}\|(\boldsymbol{P}^{\rho}(\prod_{i=1}^{J}\operatorname{diag}_{\rho}(\boldsymbol{\Upsilon}^{i})-\prod_{i=1}^{J}\mathcal{Q}_{\nu}^{+}(\operatorname{diag}_{\rho}(\boldsymbol{\Upsilon}^{i})))\tilde{\boldsymbol{S}}\boldsymbol{z})_{\mathcal{T}}\|_2\\ \leq&\sqrt{|\mathcal{T}|}\gamma\epsilon\big(\nu\sqrt{\frac{N\kappa}{2}}\big)^{J}\sqrt{eJ},\end{aligned}$$

with a probability of failure $p_\gamma \leq N \exp(-2c\gamma^2 m)$. □

Theorem 8.8. *For all unit s-sparse vectors $x \in \tilde{\Sigma}_s^N$, given a measurement matrix $\tilde{\boldsymbol{\Phi}} \in \mathbb{C}^{m\times N}$, with $\|\tilde{\boldsymbol{\Phi}}\boldsymbol{x}\|_\infty \leq \frac{\epsilon}{2}$, whose back-projection can be factorized by J submatrices whose line are κ-sparse, where all components of these sub-matrices are bounded by $\frac{\nu}{2}$. Furthermore, the matrix $\tilde{\boldsymbol{\Phi}}$ follows the (ℓ_2, ℓ_2)-RIP; the 1-bit quantized version of this factorized back-projection, following the structure defined in (8.7), is denoted by $\tilde{\boldsymbol{\Psi}}^H$. For all support set $\mathcal{T}$ such that $|\mathcal{T}| \leq 2s$. It can be shown that, with a probability of failure exceeding $N \exp(-\frac{c}{2}\gamma^2 m)$,*

$$\frac{1}{m}\|((\tilde{\boldsymbol{\Phi}}^H - \tilde{\boldsymbol{\Psi}}^H)\tilde{\boldsymbol{\Phi}}\boldsymbol{x})_\mathcal{T}\| \leq \sqrt{8seJ}\epsilon B\gamma,$$

Provided $m \geq \frac{2s}{c}\gamma^{-2}\big(\log(\frac{eN}{s}) + 2\log(1+\frac{2}{\alpha})\big)$ and with $\alpha = \eta\frac{\kappa}{4\sqrt{s}(\frac{\kappa\nu}{\sqrt{2}})^J}$ being a constant dependent on the problem.

Proof. To extend Theorem 8.7 to the set of all sparse vectors $\boldsymbol{x} \in \tilde{\Sigma}_s^N$, we follow the same step as in Section 8.4 by defining a α-covering $\mathcal{J}_\alpha$. For linear measurements, we extend Theorem 8.7 to all $\boldsymbol{u} \in \mathcal{J}_\alpha$ with a union bound. This gives us the following upper-bound on the ℓ_2 degration

$$\frac{1}{m}\|((\tilde{\boldsymbol{\Phi}}^H - \tilde{\boldsymbol{\Psi}}^H)\tilde{\boldsymbol{\Phi}}\boldsymbol{x})_\mathcal{T}\| \leq \frac{1}{m}\|((\tilde{\boldsymbol{\Phi}}^H - \tilde{\boldsymbol{\Psi}}^H)\tilde{\boldsymbol{\Phi}}\boldsymbol{u})_\mathcal{T}\| + \frac{1}{m}\|((\tilde{\boldsymbol{\Phi}}^H - \tilde{\boldsymbol{\Psi}}^H)\tilde{\boldsymbol{\Phi}}\boldsymbol{r})_\mathcal{T}\|$$

$$\leq \sqrt{2seJ}\gamma\epsilon B + \frac{1}{m}\|((\tilde{\boldsymbol{\Phi}}^H - \tilde{\boldsymbol{\Psi}}^H)\tilde{\boldsymbol{\Phi}}\boldsymbol{r})_\mathcal{T}\|, \qquad (8.16)$$

with a probability of failure $\mathbb{P} \leq |\mathcal{J}_\alpha| p_\gamma$.

The second term of (8.16), which thanks to the RIP property of $\tilde{\boldsymbol{\Phi}}$ and the properties of the α-covering see Chapter 7, can be upperbounded, for

all $\boldsymbol{x}$, by:

$$\begin{aligned}\frac{1}{m}\|((\tilde{\boldsymbol{\Phi}}^H - \tilde{\boldsymbol{\Psi}}^H)\tilde{\boldsymbol{\Phi}}\boldsymbol{r})_{\mathcal{T}}\| &\leq \frac{\sqrt{|\mathcal{T}|}}{m}\max_i |\langle(\tilde{\boldsymbol{\Phi}}^H - \tilde{\boldsymbol{\Psi}}^H)_i, \tilde{\boldsymbol{\Phi}}\boldsymbol{r}\rangle| \\ &\leq \frac{\sqrt{|\mathcal{T}|}}{m}\max_i \|(\tilde{\boldsymbol{\Phi}}^H - \tilde{\boldsymbol{\Psi}}^H)_i\|\|\tilde{\boldsymbol{\Phi}}\boldsymbol{r}\|_2 \\ &\leq 2\sqrt{s}\|\tilde{\boldsymbol{\Phi}}^H - \tilde{\boldsymbol{\Psi}}^H\|_{\infty,\infty}\alpha,\end{aligned}$$

where $\|\tilde{\boldsymbol{\Phi}}\boldsymbol{r}\|_2$ is upperbounded by $\sqrt{2m}\alpha$, and $\|(\boldsymbol{\Phi}^H - \boldsymbol{\Psi}^H)_i\| \leq \sqrt{m}\|\boldsymbol{\Phi}^H - \boldsymbol{\Psi}^H\|_{\infty,\infty}$. The ℓ_∞ norm remains to be estimated, using the fact that $\forall i \in [N], \forall j \in [m]$, then $|\boldsymbol{\Phi}_{i,j}| < |\boldsymbol{\Psi}_{i,j}|$ and Lemma 8.6, one can directly show that

$$\|\tilde{\boldsymbol{\Phi}}^H - \tilde{\boldsymbol{\Psi}}^H\|_{\infty,\infty} \leq 2\|\tilde{\boldsymbol{\Psi}}^H\|_{\infty,\infty} \leq 2\|\boldsymbol{\Psi}^H\|_{\infty,\infty} \leq 2\kappa^{-1}(\frac{\nu\kappa}{\sqrt{2}})^J.$$

The second term in (8.16) then becomes:

$$\frac{1}{m}\|((\tilde{\boldsymbol{\Phi}}^H - \tilde{\boldsymbol{\Psi}}^H)\tilde{\boldsymbol{\Phi}}\boldsymbol{r})_{\mathcal{T}}\| \leq 4\sqrt{s}(\frac{\kappa\nu}{\sqrt{2}})^J\kappa^{-1}\alpha.$$

Setting the radius of the covering as $\alpha = \gamma\sqrt{\frac{eJ}{8}}\epsilon N^{\frac{J}{2}}\kappa^{-\frac{J}{2}-1}$, finally:

$$\frac{1}{m}\|((\tilde{\boldsymbol{\Phi}}^H - \tilde{\boldsymbol{\Psi}}^H)\tilde{\boldsymbol{\Phi}}\boldsymbol{x})_{\mathcal{T}}\| \leq \sqrt{8seJ}\epsilon B\gamma,$$

with a probability of failure upper-bounded by

$$p_\gamma \leq (\frac{eN}{s})^s(1 + \frac{2}{\alpha})^{2s}N\exp(-c\gamma^2 m).$$

Which can be upper-bounded by $N\exp(-\frac{c}{2}\gamma^2 m)$, provided that

$$m \geq \frac{2s}{c}\gamma^{-2}\big(\log(\frac{eN}{s}) + 2\log(1 + \frac{2}{\alpha})\big).$$

□

Theorem 8.9. *For all unit s-sparse vectors $x \in \tilde{\Sigma}_s^N$, given a measurement*

matrix $\tilde{\boldsymbol{\Phi}} \in \mathbb{C}^{m\times N}$*, with* $\|\tilde{\boldsymbol{\Phi}}\boldsymbol{x}\|_\infty \leq \frac{\epsilon}{2}$*, whose back-projection can be factorized by* J *submatrices whose line are* κ*-sparse, where all components of these sub-matrices are bounded by* $\frac{\nu}{2}$*. Furthermore, the matrix* $\tilde{\boldsymbol{\Phi}}$ *follows the* (ℓ_2,ℓ_2)*-RIP; the 1-bit quantized version of this factorized back-projection, following the structure defined in* (8.7)*, is denoted by* $\tilde{\boldsymbol{\Psi}}^H$*. For all support set* $\mathcal{T}$ *such that* $|\mathcal{T}| \leq 2s$*. With probability exceeding* $2N\exp\left(-\frac{c}{2}m\gamma^2\right)$ *with it can be shown that :*

$$\frac{1}{m}\|((\tilde{\boldsymbol{\Phi}}^H - \tilde{\boldsymbol{\Psi}}^H)\mathcal{Q}_\epsilon^+(\tilde{\boldsymbol{\Phi}}\boldsymbol{x}))_\mathcal{T}\| \leq \sqrt{2s}\epsilon\gamma(\sqrt{eJ}B + \frac{8}{3\kappa}c(\frac{\nu\kappa}{\sqrt{2}})^J)$$

provided $m \geq \frac{2s}{c}\gamma^{-2}(\log(\frac{eN}{s}) + 2(1+\frac{2}{\alpha}))$*, with* $\alpha = \sqrt{\frac{2}{9m}}c\gamma^2\epsilon$*.*

Proof. To extend Theorem 8.7 to the set of all sparse vectors $\boldsymbol{x} \in \tilde{\Sigma}_s^N$, we follow the same step as in Section 8.4 by defining a α-covering $\mathcal{J}_\alpha$. We again extend the non-uniform bound to all element of the covering $\mathcal{J}_\alpha$ with a union-bound.

For linear measurements, we extend Theorem 8.7 to all $\boldsymbol{u} \in \mathcal{J}_\alpha$ with a union bound. This gives us the following upper-bound on the ℓ_2 degration

$$\begin{aligned}\frac{1}{m}\|((\tilde{\boldsymbol{\Phi}}^H - \tilde{\boldsymbol{\Psi}}^H)\mathcal{Q}_\epsilon^+(\tilde{\boldsymbol{\Phi}}\boldsymbol{x}))_\mathcal{T}\| \\ \leq \eta + \frac{1}{m}\|((\tilde{\boldsymbol{\Phi}}^H - \tilde{\boldsymbol{\Psi}}^H)(\mathcal{Q}_\epsilon^+(\tilde{\boldsymbol{\Phi}}\boldsymbol{u}) - \mathcal{Q}_\epsilon^+(\tilde{\boldsymbol{\Phi}}\boldsymbol{x}))_\mathcal{T}\| \qquad (8.17)\end{aligned}$$

with a probability of failure $\mathbb{P} \leq |\mathcal{J}_\alpha|p_\gamma$.

Similarly to Theorem 8.3, we use Lemma 8.10 and 8.1 to bound the second term

$$\begin{aligned}&\frac{1}{m}\|((\tilde{\boldsymbol{\Phi}}^H - \tilde{\boldsymbol{\Psi}}^H)(\mathcal{Q}_\epsilon^+(\tilde{\boldsymbol{\Phi}}\boldsymbol{u}) - \mathcal{Q}_\epsilon^+(\tilde{\boldsymbol{\Phi}}\boldsymbol{x}))_\mathcal{T}\| \\ &\quad\leq \frac{1}{m}\sqrt{2s}\|\tilde{\boldsymbol{\Phi}}^H - \tilde{\boldsymbol{\Psi}}^H\|_{\infty,\infty}\|(\mathcal{Q}_\epsilon^+(\tilde{\boldsymbol{\Phi}}\boldsymbol{u}) - \mathcal{Q}_\epsilon^+(\tilde{\boldsymbol{\Phi}}\boldsymbol{x})\|_1 \\ &\quad\leq \frac{1}{m}\sqrt{2s}|\mathcal{D}|\epsilon(\frac{\nu}{\sqrt{2}}\kappa)^J\kappa^{-1},\end{aligned}$$

where we upper-bounded the ℓ_∞-norm thanks to Lemma 8.6. Leveraging again lemma 6.1 from [XJ19] as was done in the simple matrix mutiplication

case, one can upper-bound for one $\boldsymbol{u} \in \mathcal{J}_\rho$

$$\mathbb{P}\left(|\mathcal{D}| \geq \frac{\sqrt{32}m^{\frac{3}{2}}\alpha}{\epsilon}\right) \leq |\mathcal{J}_\alpha| \exp\left(-\sqrt{\frac{18}{16}}\frac{m^{\frac{3}{2}}\alpha}{\epsilon}\right). \tag{8.18}$$

Extending (8.18) to all elements of the $\mathcal{J}_\rho$ covering gives the desired upper-bound on the second term of (8.17). We finally obtain

$$\frac{1}{m}\|((\tilde{\boldsymbol{\Phi}}^H - \tilde{\boldsymbol{\Psi}}^H)\mathcal{Q}_\epsilon(\tilde{\boldsymbol{\Phi}}\boldsymbol{x})_\mathcal{T}\| \leq \sqrt{2seJ}B\gamma\epsilon + \sqrt{|\mathcal{T}|}\sqrt{32m}\alpha(\frac{\nu}{\sqrt{2}}\kappa)^J\kappa^{-1},$$

with probability of failure : $\mathbb{P} \leq |\mathcal{J}_\alpha|N\exp\big(-cm\gamma^2\big)+|\mathcal{J}_\alpha|\exp\big(-\sqrt{\frac{18}{16}}\frac{m^{\frac{3}{2}}\alpha}{\epsilon}\big)$.

Setting $\alpha = \sqrt{\frac{2}{9m}}c\gamma^2\epsilon$ and upper-bounding the probability finally gives the desired results

$$\frac{1}{m}\|((\tilde{\boldsymbol{\Phi}}^H - \tilde{\boldsymbol{\Psi}}^H)\mathcal{Q}_\epsilon^+(\tilde{\boldsymbol{\Phi}}\boldsymbol{x})_\mathcal{T}\| \leq \sqrt{2s}\epsilon\gamma(\sqrt{eJ}B + \frac{8}{3\kappa}c(\frac{\nu\kappa}{\sqrt{2}})^J),$$

with a probability of failure upper-bounded by :

$$\mathbb{P} \leq |\mathcal{J}_\alpha|2N\exp\big(-cm\gamma^2\big) \leq 2N\exp\big(-\frac{c}{2}m\gamma^2\big),$$

provided $m \geq \frac{2s}{c}\gamma^{-2}(\log(\frac{eN}{s}) + 2(1+\frac{2}{\alpha}))$. □

Part V

Conclusions & Perspectives

Chapter 9

Conclusions and Perspectives

In this dissertation, we studied and proposed different 1-bit acquisition and processing schemes aimed at lowering the hardware requirements or the data transfer of sensors. We particularly focused on one sensing modality, namely the FMCW radar as its mathematical representation allowed for an in-depth study of the effects of the quantization and its practicality enabled straightforward confirmations of the developed theory through actual experiments.

One of the conclusions of this thesis is not that 1-bit quantization should be the new default solution for FMCW radar signal processing but it is rather to show that Quantized Compressive Sensing theory can be applied in practical settings and that the issues it raises and the limitations that it attempts to solve are indeed relevant. Indeed, we saw in various chapters of this thesis that dithering the quantization operator, be it for the acquisition or the processing, provided a noticeable gain in the reconstruction as well as in the theoretical guarantees.

Furthermore, because FMCW radars can be represented as sensors that perform a Fourier transform of the scene they measure, the conclusions of this thesis can be partially translated to other sensing methods that relies on the Fourier transform (*e.g.*, Lidar, MRI, Sonar, ...).

The conclusions of the different chapters presented in this thesis as well

Part 2 : $\mathcal{Q}_\epsilon^+$
✓High quality reconstruction
✓Dithering solves artifacts
✓Theory, Simulations, measurements
? Reconstruction bound for QIHT with RIP
? How to implement the add dither ?
? More complex model, SAR ?
? Noise, and other effects

Part 4 : $\mathcal{Q}_\nu^+(\mathbf{\Phi}^H)$
✓Reconstruction bound for QPBP(Q) for direct and factorized model
✓Theory and simulations (QFFT)
? Better dependency on s, noise, metric ?
? Extension: network, Unfolding,...

Part 3
$\text{sign}_{\mathbb{C}}$
✓Reconstruction PBP with PO meas
✓l1l2 RIP for complex gaussian
? Application with $\text{sign}_{\mathbb{C}}$
? Other model with (ℓ_1, ℓ_2)-RIP
? Tighter bound ?
$\mathcal{Q}^{\odot}$
✓Efficient implementation
✓Non random dither
✓Limitations of PO CS with Fourier
✓Theory, simulations, measurements
? Prototype
? More complex model
? Stronger guarantees/algo for POCS

Table 9.1: Table summarizing the contributions and perspectives of this thesis; The cells in **blue** are the contributions; **red** cells are the perspectives and possible future works.

as possible perspectives are discussed hereafter and summarized in Table 9.1.

Part II studied the coarse quantization of radar signals using a dithered acquisition. In Chapter 4 and 5 we showed that in a noiseless case, radar measurements, when quantized to 1-bit, must be dithered in order to guarantee a successful reconstruction. Indeed, we showed in Chapter 4 that the absence of dithering before the acquisition can create ambiguous scenarios where the quantization of different sparse vectors (or range profiles) results in the identical measured bits. The extension of the model in Chapter 5 from the simple range estimation to the joint range and angle-of-arrival estimation showed that a similar issue can render the estimation of the angle of arrival impossible, even when only measuring one target. Adding a random dithering that is carefully linked to the dynamic range of the measured signal solves this issue and the performances it enables was shown through simulations and actual radar measurements. The theory developed for PBP in [XJ19] for the 1D ranging was extended in Chapter 5 for the 2D estimation. Possible perspectives are listed below

- The developed bound in Chapter 5 and in [XJ19] only applies to the PBP algorithm. A potential future work would be the extension of these bounds to the iterative QIHT algorithm for RIP matrices. The simulations showed that the reconstructions followed $\mathcal{O}(m^{-1})$, this was also proved for Gaussian matrices in [Fri+20].

- How to implement efficiently this additive dither efficiently is still an

open question. Indeed, the conditions attached to this process for it to be successful, *i.e.*, generating a random variable that follows a uniform distribution and that is scaled according to the dynamic of the measured are quite stringent. Chapter 7 tries to answer this by proposing an alternative to the additive dither.

- All the work presented in this thesis, apart from Chapter 6, assumes a noiseless settings. It is still an open question as to what is the best methodology regarding noisy 1-bit measurements. Indeed, Chapter 5 showed in Section 5.6 that in the presence of noise, using a uniform dither at *full dynamic* does not yield the best results. How to scale this dither with respect to the power of the already present noise is a challenge. In fact it mirrors another challenge of the additive dithering which is that the 1-bit quantization removes the information about the amplitude which makes the estimation of the power of the noise, as well as the dynamic of the signal complicated. A possibility would be to switch from a uniform distribution to a Gaussian one, as to better control the distribution of the dither in case of noisy measurement, as it will remain Gaussian. This method however is also limited as the behaviour of the quantizer will depart from Lemma 3.6 from $\mathbb{E}\{\mathcal{Q}_\epsilon(\boldsymbol{y}+\boldsymbol{\xi})\} = \boldsymbol{y}$ with $\boldsymbol{\xi} \sim \mathcal{U}[-\frac{\epsilon}{2}, \frac{\epsilon}{2}]$ to $\mathbb{E}\{\mathcal{Q}_\epsilon(\boldsymbol{y}+\boldsymbol{n})\} = \operatorname{erf}(\frac{\boldsymbol{y}}{\sigma\sqrt{2}})$ with $\boldsymbol{n} \sim \mathcal{N}(0, \sigma^2)$, *i.e.*, the level of the noise with respect to the signal will determine to which extent the signal can be recovered. This has been partially covered in [PV16] but only applied to measurement matrix that are Gaussian, which is of limited practical interest for more applied sensing scenarios.

- Chapter 4 and 5 only considered fairly basic model in order to clearly highlight the gain provided by the dithered acquisition. It would be of interest to now extend this study to more complex model, *e.g.*, MIMO radars, SAR, or other sensing methods like LIDARs.

Part III was dedicated to the study of the Phase-Only acquisition in Chapter 6 and another 1-bit acquisition strategy in Chapter 7, namely the 1-bit quantization with multiplicative dithering which can be seen as a quantized extension of the $\operatorname{sign}_{\mathbb{C}}$ acquisition.

In Chapter 6, we showed that the Phase Only acquisition is an extension

of the well-known 1-bit CS setting to the complex domain. The reconstruction bound of the PBP algorithm that was proven for real matrices that follow the (ℓ_1, ℓ_2)-RIP in was shown to also hold in the complex case. We also demonstrated that complex Gaussian matrices can also exhibit this property. Extensions and future work regarding this chapter are listed hereafter.

- The bound developed in Theorem 6.2 could be refined from $\mathcal{O}(m^{-\frac{1}{4}})$ to $\mathcal{O}(m^{-\frac{1}{2}})$ to match the behaviour observed in the simulations in Sec 6.6.

- Since the first publication of the results of this chapter in [Feu+19], the results were extended in [JF21] to *instance-optimal* algorithms. This work showed that the problem of recovering real sparse vector from complex phase-only measurements, can be recasted as a linear problem that can be solved by *instance-optimal* algorithms if this new model respect the RIP. This was shown for complex Gaussian measurement matrices.

- This chapter was dedicated to the special case of complex measurements matrix that follows the (ℓ_1, ℓ_2)-RIP, it could be interesting to extend it to other models.

Part II introduced the use of dithering the 1-bit measurements coming from radar signals but did not consider how to implement it. Chapter 7 first highlighted the constraints associated to the use of an additive dithering in a radar context and in response proposed the used of a multiplicative one. This acquisition method was linked to the PO-CS setting and shown to have, compared to the additive dither, less constraints on the actual implementation of the dithering process. The link with the PO-CS was studied further by showing that, similarly to the 1-bit acquisition, PO measurements can also be ambiguous for the special case of Fourier based models. We also demonstrated that the performances of the reconstruction of the multiplicative scheme can be upper-bound by the reconstruction of the PO-CS measurements to which a discrepancy term is added. This terms has been shown, through a non uniform bound, to behave as $\mathcal{O}(m^{-\frac{1}{2}})$, *i.e.*, it can be made arbitrary small for a sufficient number of measurements. We also introduced a relaxed version of the proposed multiplicative dithering

that only relied on a deterministic function that can be efficiently implemented. These results were then confirmed using extensive Monte-Carlo simulations and real radar measurements, highlighting the fact that the proposed multiplicative 1-bit acquisition is a interesting trade-off between theoretical guarantees, performances and ease of implementations. Some perspective regarding this work are provided hereafter.

- The big advantage of the proposed multiplicative dithering procedure is its ease of implementation. The next logical step is thus to implement it in an actual radar prototype. A future publications will study the performances of an actual FMCW radar that is multiplicatively dithering using the architecture in Fig. 7.3. The use of the structured and deterministic dither will also be investigated.

- The radar model considered was only the estimation of the range. A possible perspective is to extend this model to more than the range estimation by adding the angle of arrival of the velocity using the Doppler Effect.

- Lastly, this work showed that PO measurements with Fourier based model were ambiguous for complex sparse signal. While in this specific settings, this fact limits the ability to uniformly-bound the reconstruction but it could be worthwhile to study alternative properties of these coarse measurements in order to infer more about the possible reconstruction. For example, in [Jac+13], the authors departed from the ℓ_2-norm between sparse vector to use instead the angle between them to develop the Binary-ϵ-Stable-Embedding properties for 1-bit CS. A similar methodology could be interesting.

The last part of this thesis (Part IV) focused on the quantization of the matrices used for the back-projection. We studied two quantization schemes, namely the direct quantization of the back-projection operator ($\mathcal{Q}_\nu^+(\mathbf{\Phi}^H)$) and the quantization of the coefficients of the sparse and factorized representation of matrices like the FFT. Lowering the resolution of these operators can simplify the hardware implementation of the arithmetical function used by the processing. For both quantization procedure, we showed a uniform reconstruction guarantee using this modified PBP (*i.e.*, QPBP) and proved that it follows the classic bound of PBP $\mathcal{O}(m^{-\frac{1}{2}})$ up to

a constant. The quality of the developed bounds were then evaluated using extensive Monte Carlo simulations. Perspectives on this topic are provided below.

- The reconstruction bounds developed in this work (*i.e.*, Theorem 8.3, Theorem 8.2, Theorem 8.8 and Theorem 8.9) all exhibit an increased in the dependency on the sparsity of the vector. Indeed, classic reconstruction bound in compressive sensing follow $\mathcal{O}(\sqrt{s})$ instead of the $\mathcal{O}(s)$. It could be interesting to investigate if this bound be tightened to match the high-resolution bound (up to a constant).

- It is interesting to note that Theorem 8.7 and 8.8 can be used to extend the reconstruction results from the noiseless case (*i.e.*, $\boldsymbol{z} = \boldsymbol{\Phi}\boldsymbol{x}$) to a noisy one (*i.e.*, $\boldsymbol{z} = \boldsymbol{\Phi}\boldsymbol{x} + \boldsymbol{n}$) where $\boldsymbol{n} \in \mathbb{C}^m$ is ℓ_∞-bounded noise. Indeed, the Quantized BP of noisy linear measurements can be upper-bound by

$$\begin{aligned}\|\boldsymbol{x} - \mathsf{H}_s(\tilde{\boldsymbol{\Psi}}^H \boldsymbol{z})\|_2 \leq \|x - (\tilde{\boldsymbol{\Phi}}^H \boldsymbol{z})_{\mathcal{T}}\|_2 \\ + \|\big((\tilde{\boldsymbol{\Psi}}^H - \tilde{\boldsymbol{\Phi}}^H)\tilde{\boldsymbol{\Phi}}\boldsymbol{x}\big)_{\mathcal{T}}\|_2 + \|\big((\tilde{\boldsymbol{\Psi}}^H - \tilde{\boldsymbol{\Phi}}^H)\boldsymbol{n}\big)_{\mathcal{T}}\|_2, \quad (9.1)\end{aligned}$$

  where $\mathcal{T} = \operatorname{supp}(\boldsymbol{x}) \cup \operatorname{supp}(\mathsf{H}_s(\tilde{\boldsymbol{\Psi}}^H \boldsymbol{z}))$. From (9.1) it is clear that in order to bound the reconstruction, one needs to add to the bound developed in Theorem 8.8 the non-uniform bound on the discrepancy (*i.e.*, Theorem 8.7) but not on the measurements but on the bounded noise $\boldsymbol{n}$. From this simple fact, it is clear that extending the analysis presented in Chapter 8 from the noiseless to a noisy case should be investigated.

- The simulations results presented in Section 8.6 finished with a comparison between the direct quantization and the factorized one. The metric used to compare their reconstruction performances was the complexity attached to the application of these operators (*i.e.*, $\mathcal{O}(mN)$ and $\mathcal{O}(m\log(N))$). This metric however is imperfect and this comparison should be furthered using a more in-depth measure of the complexity of these scheme using, for example as a benchmark, a specific platform on which signal processing method can be implemented.

- The second part focused on back-projection that can be factorized

as multiplication of different sub-matrices. It could be interesting to extend this to other algorithms whose estimation process can be expressed as the application of matrices, for example Neural Networks [Mer+16] and other algorithms related to Unfolding [MLE21].

Bibliography

[AMG] https://www.amg-microwave.com/-Test-et-Mesure- AMG target simulator AMG-043-007. *KMD2 radar transceiver*.

[Bar+08] Richard G. Baraniuk et al. "A Simple Proof of the Restricted Isometry Property for Random Matrices". In: *Constructive Approximation* 28 (2008), pp. 253–263.

[Bar+17] Richard Baraniuk et al. "Exponential Decay of Reconstruction Error From Binary Measurements of Sparse Signals". In: *IEEE Transactions on Information Theory* 63 (2017), pp. 3368–3385.

[BB08] P. T. Boufounos and R. G. Baraniuk. "1-Bit compressive sensing". In: *2008 42nd Annual Conference on Information Sciences and Systems*. 2008, pp. 16–21. DOI: 10.1109/CISS.2008.4558487.

[BD09] Thomas Blumensath and Mike E. Davies. "Iterative hard thresholding for compressed sensing". In: *Applied and Computational Harmonic Analysis* 27.3 (2009), pp. 265–274. ISSN: 1063-5203. DOI: https://doi.org/10.1016/j.acha.2009.04.002. URL: https://www.sciencedirect.com/science/article/pii/S1063520309000384.

[Bha+11] Subhankar Bhattacharjee et al. "Evaluation of power efficient adder and multiplier circuits for FPGA based DSP applications". In: *2011 International Conference on Communication and Industrial Application*. 2011, pp. 1–5. DOI: 10.1109/ICCIndA.2011.6146691.

[Bil+19] Igal Bilik et al. “The Rise of Radar for Autonomous Vehicles: Signal Processing Solutions and Future Research Directions”. In: *IEEE Signal Processing Magazine* 36.5 (2019), pp. 20–31. DOI: 10.1109/MSP.2019.2926573.

[Bou+15] Petros T. Boufounos et al. “Quantization and Compressive Sensing”. In: *Compressed Sensing and its Applications: MATHEON Workshop 2013*. Ed. by Holger Boche et al. Cham: Springer International Publishing, 2015, pp. 193–237. ISBN: 978-3-319-16042-9. DOI: 10.1007/978-3-319-16042-9_7. URL: https://doi.org/10.1007/978-3-319-16042-9_7.

[Bou12] P. T. Boufounos. “Universal Rate-Efficient Scalar Quantization”. In: *IEEE Transactions on Information Theory* 58.3 (2012), pp. 1861–1872. DOI: 10.1109/TIT.2011.2173899.

[Bou13a] Petros T Boufounos. “Sparse Signal Reconstruction from Phase-only Measurements”. In: *10th international conference on Sampling Theory and Applications (SampTA 2013)*. Bremen, Germany, July 2013, pp. 256–259.

[Bou13b] Petros T. Boufounos. “Angle-preserving quantized phase embeddings”. In: ed. by Dimitri Van De Ville, Vivek K. Goyal, and Manos Papadakis. Vol. 8858. International Society for Optics and Photonics. SPIE, 2013, pp. 375–383. DOI: 10.1117/12.2024412. URL: https://doi.org/10.1117/12.2024412.

[Bra96] B. Brannon. “a Overcoming Converter Nonlinearities with Dither by Brad Brannon AN-410 APPLICATION NOTE”. In: 1996.

[Bru+11] Stephan Brusch et al. “Ship Surveillance With TerraSAR-X”. In: *IEEE Transactions on Geoscience and Remote Sensing* 49.3 (2011), pp. 1092–1103. DOI: 10.1109/TGRS.2010.2071879.

[BS07] R. Baraniuk and P. Steeghs. “Compressive Radar Imaging”. In: *2007 IEEE Radar Conference*. 2007, pp. 128–133. DOI: 10.1109/RADAR.2007.374203.

[Buh+98] J.M. Buhmann et al. “Dithered Color Quantization”. In: *Computer Graphics Forum* 17.3 (1998), pp. 219–231. DOI: https://doi.org/10.1111/1467-8659.00269. eprint: https:

//onlinelibrary.wiley.com/doi/pdf/10.1111/1467-8659.00269. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-8659.00269.

[CA04] L. Cheded and S. Akhtar. "A novel and fast 1-bit FFT scheme with two dither-quantized channels". In: *2004 12th European Signal Processing Conference*. Sept. 2004, pp. 1521–1524.

[CE18] D. Cohen and Y. C. Eldar. "Sub-Nyquist Radar Systems: Temporal, Spectral, and Spatial Compression". In: *IEEE Signal Processing Magazine* 35.6 (2018), pp. 35–58. DOI: 10.1109/MSP.2018.2868137.

[CRT06a] E. J. Candes, J. Romberg, and T. Tao. "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information". In: *IEEE Transactions on Information Theory* 52.2 (2006), pp. 489–509. DOI: 10.1109/TIT.2005.862083.

[CRT06b] Emmanuel J. Candes, Justin K. Romberg, and Terence Tao. "Stable signal recovery from incomplete and inaccurate measurements". In: *Communications on Pure and Applied Mathematics* 59.8 (2006), pp. 1207–1223. DOI: https://doi.org/10.1002/cpa.20124. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpa.20124. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.20124.

[CT05] E. J. Candes and T. Tao. "Decoding by linear programming". In: *IEEE Transactions on Information Theory* 51.12 (2005), pp. 4203–4215. DOI: 10.1109/TIT.2005.858979.

[CT65] James W. Cooley and John W. Tukey. "An Algorithm for the Machine Calculation of Complex Fourier Series". In: *Mathematics of Computation* 19.90 (1965).

[Dar+19] Davide Dardari et al. "An Ultra-wideband Battery-less Positioning System for Space Applications". In: *2019 IEEE International Conference on RFID Technology and Applications (RFID-TA)*. 2019, pp. 104–109. DOI: 10.1109/RFID-TA.2019.8892114.

[Dar+20] Davide Dardari et al. “An Ultra-Low Power Ultra-Wide Bandwidth Positioning System”. In: *IEEE Journal of Radio Frequency Identification* 4.4 (2020), pp. 353–364. DOI: 10.1109/JRFID.2020.3008200.

[Dir19] Sjoerd Dirksen. “Quantized Compressed Sensing: A Survey”. In: *Compressed Sensing and Its Applications: Third International MATHEON Conference 2017*. Ed. by Holger Boche et al. Cham: Springer International Publishing, 2019, pp. 67–95. ISBN: 978-3-319-73074-5. DOI: 10.1007/978-3-319-73074-5_2. URL: https://doi.org/10.1007/978-3-319-73074-5_2.

[DJR19] Sjoerd Dirksen, Hans Christian Jung, and Holger Rauhut. “One-bit compressed sensing with partial Gaussian circulant matrices”. In: *Information and Inference: A Journal of the IMA* 9.3 (Oct. 2019), pp. 601–626. ISSN: 2049-8772. DOI: 10.1093/imaiai/iaz017. eprint: https://academic.oup.com/imaiai/article-pdf/9/3/601/33720745/iaz017.pdf. URL: https://doi.org/10.1093/imaiai/iaz017.

[Don06a] D. L. Donoho. “Compressed sensing”. In: *IEEE Transactions on Information Theory* 52.4 (2006), pp. 1289–1306. DOI: 10.1109/TIT.2006.871582.

[Don06b] David L. Donoho. “For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution”. In: *Communications on Pure and Applied Mathematics* 59.6 (2006), pp. 797–829. DOI: 10.1002/cpa.20132. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpa.20132. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.20132.

[DZ15] X. Dong and Y. Zhang. “A MAP Approach for 1-Bit Compressive Sensing in Synthetic Aperture Radar Imaging”. In: *IEEE Geoscience and Remote Sensing Letters* 12.6 (2015), pp. 1237–1241. DOI: 10.1109/LGRS.2015.2390623.

[EL04] Miloš D. Ercegovac and Tomás Lang. “CHAPTER 4 - Multiplication”. In: *Digital Arithmetic*. Ed. by Miloš D. Ercegovac and Tomás Lang. The Morgan Kaufmann Series in Computer Architecture and Design. San Francisco: Morgan Kaufmann, 2004, pp. 180–245. DOI: https://doi.org/10.1016/B978-155860798-9/50006-3. URL: https://www.sciencedirect.com/science/article/pii/B9781558607989500063.

[End13] J. Ender. “A brief review of compressive sensing applied to radar”. In: *2013 14th International Radar Symposium (IRS)*. Vol. 1. 2013, pp. 3–16.

[Feu+17] Thomas Feuillen et al. “Localization of Rotating Targets Using a Monochromatic Continuous-Wave Radar”. In: *IEEE Antennas and Wireless Propagation Letters* 16 (2017), pp. 2598–2601. DOI: 10.1109/LAWP.2017.2735020.

[Feu+18a] Thomas Feuillen et al. “1-bit Localization Scheme for Radar using Dithered Quantized Compressed Sensing”. In: *2018 5th International Workshop on Compressed Sensing applied to Radar, Multimodal Sensing, and Imaging (CoSeRa)* (2018).

[Feu+18b] Thomas Feuillen et al. “Quantity over Quality: Dithered Quantization for Compressive Radar Systems”. In: *2019 IEEE Radar Conference (RadarConf)* (2018), pp. 1–6.

[Feu+19] Thomas Feuillen et al. *(l1,l2)-RIP and Projected Back-Projection Reconstruction for Phase-Only Measurements*. 2019. arXiv: 1912.02880 [eess.SP].

[Feu+20] Thomas Feuillen et al. *One Bit to Rule Them All : Binarizing the Reconstruction in 1-bit Compressive Sensing*. 2020. arXiv: 2008.07264 [eess.SP].

[FL18] S. Foucart and Jiangyuan Li. “Sparse Recovery from Inaccurate Saturated Measurements”. In: *Acta Applicandae Mathematicae* 158 (2018), pp. 49–66.

[FMV16] Thomas Feuillen, Achraf Mallat, and Luc Vandendorpe. “Stepped frequency radar for automotive application: Range-Doppler coupling and distortions analysis”. In: *MILCOM 2016 - 2016 IEEE*

Military Communications Conference. 2016, pp. 894–899. DOI: 10.1109/MILCOM.2016.7795443.

[Fou17] Simon Foucart. "Flavors of Compressive Sensing". In: *Approximation Theory XV: San Antonio 2016*. Ed. by Gregory E. Fasshauer and Larry L. Schumaker. Cham: Springer International Publishing, 2017, pp. 61–104. ISBN: 978-3-319-59912-0.

[FR13] Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing*. Birkhäuser Basel, 2013. ISBN: 0817649476.

[Fri+20] M. Friedlander et al. "NBIHT: An Efficient Algorithm for 1-bit Compressed Sensing with Optimal Error Decay Rate". In: *ArXiv* abs/2012.12886 (2020).

[FVJ18] Thomas Feuillen, Luc Vandendorpe, and Laurent Jacques. *An extreme bit-rate reduction scheme for 2D radar localization*. 2018. arXiv: 1812.05359 [eess.SP].

[Gal+19] Gilles Monnoyer de Galland de Carnieres et al. "Sparsity-Driven Moving Target Detection in Distributed Multistatic FMCW Radars". In: *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*. 2019, pp. 151–155. DOI: 10.1109/CAMSAP45676.2019.9022656.

[Gal+20] Gilles Monnoyer de Galland et al. *Going Below and Beyond, Off-the-Grid Velocity Estimation from 1-bit Radar Measurements*. 2020. arXiv: 2011.05034 [eess.SP].

[Gal+21] Gilles Monnoyer de Galland et al. *Sparse Factorization-based Detection of Off-the-Grid Moving targets using FMCW radars*. 2021. arXiv: 2102.05072 [eess.SP].

[GS93] R. M. Gray and T. G. Stockham. "Dithered quantizers". In: *IEEE Transactions on Information Theory* 39.3 (1993), pp. 805–812. DOI: 10.1109/18.256489.

[GTH71] J. Graeme, G. E. Tobey, and L. Huelsman. "Operational Amplifiers; Design and Applications". In: 1971.

[Gun+10] C. Gunturk et al. “Sobolev Duals for Random Frames and Sigma-Delta Quantization of Compressed Sensing Measurements”. In: *CoRR* abs/1002.0182 (Jan. 2010).

[Gün+10] S. Güntürk et al. “Sobolev Duals for Random Frames and Sigma-Delta Quantization of Compressed Sensing Measurements”. In: *arXiv e-prints*, arXiv:1002.0182 (Feb. 2010), arXiv:1002.0182. arXiv: 1002.0182 [cs.IT].

[Haf+20] Stephan Hafner et al. “Mitigation of Leakage in FMCW Radars by Background Subtraction and Whitening”. In: *IEEE Microwave and Wireless Components Letters* 30.11 (2020), pp. 1105–1107. DOI: 10.1109/LMWC.2020.3023850.

[HN10] Matthew A. Herman and Deanna Needell. *Mixed Operators in Compressed Sensing*. 2010. arXiv: 1004.0033 [math.NA].

[HS09] M. A. Herman and T. Strohmer. “High-Resolution Radar via Compressed Sensing”. In: *IEEE Transactions on Signal Processing* 57.6 (2009), pp. 2275–2284. DOI: 10.1109/TSP.2009.2014277.

[HS10] Matthew A. Herman and Thomas Strohmer. “General Deviants: An Analysis of Perturbations in Compressed Sensing”. In: *IEEE Journal of Selected Topics in Signal Processing* 4 (2010), pp. 342–349.

[Hua+20] De Huang et al. “Matrix Concentration for Products”. In: (2020). arXiv: 2003.05437 [math.PR].

[Hul06] Christian Hulsmeyer. *Wireless Transmitting and Receiving Mechanism for Electric Waves*. U.S. Patent 1810 150, Jan. 1906.

[Jac+13] L. Jacques et al. “Robust 1-Bit Compressive Sensing via Binary Stable Embeddings of Sparse Vectors”. In: *IEEE Transactions on Information Theory* 59.4 (2013), pp. 2082–2102. DOI: 10.1109/TIT.2012.2234823.

[Jac16] L. Jacques. “Error Decay of (Almost) Consistent Signal Estimations From Quantized Gaussian Random Projections”. In: *IEEE Transactions on Information Theory* 62.8 (2016), pp. 4696–4709. DOI: 10.1109/TIT.2016.2573313.

[JC17] Laurent Jacques and Valerio Cambareri. "Time for dithering: fast and quantized random embeddings via the restricted isometry property". In: *Information and Inference: A Journal of the IMA* 6.4 (Apr. 2017), pp. 441–476. ISSN: 2049-8764. DOI: 10.1093/imaiai/iax004. eprint: https://academic.oup.com/imaiai/article-pdf/6/4/441/22874344/iax004.pdf. URL: https://doi.org/10.1093/imaiai/iax004.

[JDD13] Laurent Jacques, Kévin Degraux, and Christophe De Vleeschouwer. "Quantized Iterative Hard Thresholding: Bridging 1bit and High-Resolution Quantized Compressed Sensing". In: *10th international conference on Sampling Theory and Applications (SampTA 2013)*. Bremen, Germany, July 2013, pp. 105–108.

[JF21] Laurent Jacques and Thomas Feuillen. *The importance of phase in complex compressive sensing*. 2021. DOI: 10.1109/TIT.2021.3073566.

[JHF11] L. Jacques, D. K. Hammond, and J. M. Fadili. "Dequantizing Compressed Sensing: When Oversampling and Non-Gaussian Constraints Combine". In: *IEEE Transactions on Information Theory* 57.1 (2011), pp. 559–571. DOI: 10.1109/TIT.2010.2093310.

[Kam+12] U. Kamilov et al. "One-Bit Measurements With Adaptive Thresholds". In: *IEEE Signal Processing Letters* 19 (2012), pp. 607–610.

[Kay93] S.M. Kay. *Fundamentals of Statistical Signal Processing: Detection theory*. Fundamentals of Statistical Signal Processing. PTR Prentice-Hall, 1993. ISBN: 9780133457117. URL: https://books.google.nl/books?id=aFwESQAACAAJ.

[Las+11] Jason N. Laska et al. "Democracy in action: Quantization, saturation, and compressive sensing". In: *Applied and Computational Harmonic Analysis* 31.3 (2011), pp. 429–443. ISSN: 1063-5203. DOI: https://doi.org/10.1016/j.acha.2011.02.002. URL: http://www.sciencedirect.com/science/article/pii/S1063520311000248.

[LDP07] Michael Lustig, David Donoho, and John M. Pauly. "Sparse MRI: The application of compressed sensing for rapid MR imaging". In: *Magnetic Resonance in Medicine* 58.6 (2007), pp. 1182–1195. DOI: https://doi.org/10.1002/mrm.21391. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/mrm.21391. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.21391.

[Le+05] Bin Le et al. "Analog-to-digital converters". In: *IEEE Signal Processing Magazine* 22.6 (2005), pp. 69–77. DOI: 10.1109/MSP.2005.1550190.

[Led91] Michel Ledoux. *Probability in Banach Spaces : Isoperimetry and Processes*. Berlin, Heidelberg: Springer Berlin Heidelberg, 1991. ISBN: 978-3-642-20211-7.

[Li+16] J. Li et al. "Compressive radar sensing via one-bit sampling with time-varying thresholds". In: *2016 50th Asilomar Conference on Signals, Systems and Computers*. 2016, pp. 1164–1168. DOI: 10.1109/ACSSC.2016.7869554.

[Llo52] E. H. Lloyd. "Least-Squares Estimation of Location and Scale Parameters Using Order Statistics". In: *Biometrika* 39.1/2 (1952), pp. 88–95. ISSN: 00063444. URL: http://www.jstor.org/stable/2332466.

[Mer+16] P. Merolla et al. "Deep neural networks are robust to weight binarization and other non-linear distortions". In: *ArXiv* abs/1606.01981 (2016).

[MLE21] Vishal Monga, Yuelong Li, and Yonina C. Eldar. "Algorithm Unrolling: Interpretable, Efficient Deep Learning for Signal and Image Processing". In: *IEEE Signal Processing Magazine* 38.2 (2021), pp. 18–44. DOI: 10.1109/MSP.2020.3016905.

[MSM17] Bernard Meyer, Johann B. de Swardt, and Paul van der Merwe. "Comparing baseband and intermediate frequency FMCW radar receivers". In: *2017 IEEE AFRICON*. 2017, pp. 563–568. DOI: 10.1109/AFRCON.2017.8095543.

[NT09] D. Needell and J.A. Tropp. "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples". In: *Applied and Computational Harmonic Analysis* 26.3 (2009), pp. 301–321. ISSN: 1063-5203. DOI: https://doi.org/10.1016/j.acha.2008.07.002. URL: https://www.sciencedirect.com/science/article/pii/S1063520308000638.

[OHL82] Alan V. Oppenheim, Monson H. Hayes, and Jae S. Lim. "Iterative Procedures For Signal Reconstruction From Fourier Transform Phase". In: *Optical Engineering* 21.1 (1982), pp. 122–127. DOI: 10.1117/12.7972871. URL: https://doi.org/10.1117/12.7972871.

[OL81] A. V. Oppenheim and J. S. Lim. "The importance of phase in signals". In: *Proceedings of the IEEE* 69.5 (May 1981), pp. 529–541. ISSN: 1558-2256. DOI: 10.1109/PROC.1981.12022.

[Pap02] Athanasios Papoulis. *Probability, random variables, and stochastic processes*. Boston: McGraw-Hill, 2002. ISBN: 0071226613.

[PCS11] Jason T. Parker, Volkan Cevher, and Philip Schniter. "Compressive sensing under matrix uncertainties: An Approximate Message Passing approach". In: *2011 Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*. 2011, pp. 804–808. DOI: 10.1109/ACSSC.2011.6190118.

[PV12] Yaniv Plan and Roman Vershynin. *One-bit compressed sensing by linear programming*. 2012. arXiv: 1109.4299 [cs.IT].

[PV13] Y. Plan and R. Vershynin. "Robust 1-bit Compressed Sensing and Sparse Logistic Regression: A Convex Programming Approach". In: *IEEE Transactions on Information Theory* 59 (2013), pp. 482–494.

[PV14] Yaniv Plan and Roman Vershynin. "Dimension Reduction by Random Hyperplane Tessellations". In: *Discrete & Computational Geometry* 51.2 (Mar. 2014), pp. 438–461. ISSN: 1432-0444. DOI: 10.1007/s00454-013-9561-6. URL: https://doi.org/10.1007/s00454-013-9561-6.

[PV16] Y. Plan and R. Vershynin. “The Generalized Lasso With Non-Linear Observations”. In: *IEEE Transactions on Information Theory* 62.3 (2016), pp. 1528–1537. DOI: 10.1109/TIT.2016.2517008.

[Rau10] Holger Rauhut. “Compressive Sensing and Structured Random Matrices”. In: *Theoretical Foundations and Numerical Methods for Sparse Recovery*. Ed. by Massimo Fornasier. De Gruyter, 2010, pp. 1–92. DOI: doi:10.1515/9783110226157.1. URL: https://doi.org/10.1515/9783110226157.1.

[RDD18] Meenu Rani, S. B. Dhok, and R. B. Deshmukh. “A Systematic Review of Compressive Sensing: Concepts, Implementations and Applications”. In: *IEEE Access* 6 (2018), pp. 4875–4894. DOI: 10.1109/ACCESS.2018.2793851.

[RDG18] J. Rapp, R. M. A. Dawson, and V. K. Goyal. “Improving Lidar Depth Resolution with Dither”. In: *2018 25th IEEE International Conference on Image Processing (ICIP)*. 2018, pp. 1553–1557. DOI: 10.1109/ICIP.2018.8451528.

[RFB] https://www.rfbeam.ch/product?id=21 RFBeam. *KMD2 radar transceiver*.

[Sha49] C. E. Shannon. “Communication in the Presence of Noise”. In: *Proceedings of the IRE* 37.1 (1949), pp. 10–21. DOI: 10.1109/JRPROC.1949.232969.

[Sko80] M.I. Skolnik. *Introduction to Radar Systems /2nd Edition/*. 2nd ed. New York: McGraw Hill Book Co., 1980.

[SSS06] Stefan Scheiblhofer, Stefan Schuster, and Andreas Stelzer. “Effects of Systematic FMCW Radar Sweep Nonlinearity on Bias and Variance of Target Range Estimation”. In: *2006 IEEE MTT-S International Microwave Symposium Digest*. 2006, pp. 1418–1421. DOI: 10.1109/MWSYM.2006.249535.

[Sta+16] Stephan Stanko et al. “Millimeter resolution SAR imaging of infrastructure in the lower THz region using MIRANDA-300”. In: *2016 46th European Microwave Conference (EuMC)*. 2016, pp. 1505–1508. DOI: 10.1109/EuMC.2016.7824641.

[Ver18] Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018. DOI: 10.1017/9781108231596.

[Vin+11] Gabor Vinci et al. "An ultrawideband antenna for FMCW-radar positioning systems". In: *Proceedings of the 5th European Conference on Antennas and Propagation (EUCAP)*. 2011, pp. 372–374.

[Wal99] R.H. Walden. "Analog-to-digital converter survey and analysis". In: *IEEE Journal on Selected Areas in Communications* 17.4 (1999), pp. 539–550. DOI: 10.1109/49.761034.

[Wan+16] Saiwen Wang et al. "Interacting with Soli: Exploring Fine-Grained Dynamic Gesture Recognition in the Radio-Frequency Spectrum". In: *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. UIST '16. Tokyo, Japan: Association for Computing Machinery, 2016, pp. 851–860. ISBN: 9781450341899. DOI: 10.1145/2984511.2984565. URL: https://doi.org/10.1145/2984511.2984565.

[Wan+19] X. Wang et al. "Enhanced 1-Bit Radar Imaging by Exploiting Two-Level Block Sparsity". In: *IEEE Transactions on Geoscience and Remote Sensing* 57.2 (2019), pp. 1131–1141. DOI: 10.1109/TGRS.2018.2864795.

[Wan97] R. Wannamaker. "The Theory of Dithered Quantization". In: 1997.

[Wat+18] Manabu Watanabe et al. "Early-Stage Deforestation Detection in the Tropics With L-band SAR". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 11.6 (2018), pp. 2127–2133. DOI: 10.1109/JSTARS.2018.2810857.

[XJ19] Chunlei Xu and Laurent Jacques. "Quantized compressive sensing with RIP matrices: the benefit of dithering". In: *Information and Inference: A Journal of the IMA* 9.3 (Nov. 2019), pp. 543–586. ISSN: 2049-8772. DOI: 10.1093/imaiai/iaz021. eprint: https://academic.oup.com/imaiai/article-pdf/

9/3/543/33720723/iaz021.pdf. URL: https://doi.org/10.1093/imaiai/iaz021.

[XSJ18] C. Xu, V. Schellekens, and L. Jacques. "Taking the Edge off Quantization: Projected Back Projection in Dithered Compressive Sensing". In: *2018 IEEE Statistical Signal Processing Workshop (SSP)*. 2018, pp. 203–207. DOI: 10.1109/SSP.2018.8450784.

[ZLT12] Jianmin Zhou, Zhen Li, and Panpan Tang. "Mountain glacier motion change detection by satellite L-band SAR data". In: *2012 IEEE International Geoscience and Remote Sensing Symposium*. 2012, pp. 4434–4437. DOI: 10.1109/IGARSS.2012.6350488.

[ZO18] Seyed Alireza Zahrai and Marvin Onabajo. "Review of Analog-To-Digital Conversion Characteristics and Design Considerations for the Creation of Power-Efficient Hybrid Data Converters". In: *Journal of Low Power Electronics and Applications* 8.2 (2018). ISSN: 2079-9268. DOI: 10.3390/jlpea8020012. URL: https://www.mdpi.com/2079-9268/8/2/12.