THE UNIVERSITY
of EDINBURGH

Compressive Learning: New Models and Applications

Michael Patrick Sheehan

A thesis submitted for the degree of Doctor of Philosophy.
The University of Edinburgh.
December 2021

Abstract

Today's world is fuelled by data. From self-driving cars through to agriculture, massive amounts of data are used to fit learning models to provide valuable insights and predictions. Such insights come at a significant price as many traditional learning procedures have both memory and computational costs that scale with the size of the data. This quickly becomes prohibitive, even when substantial resources are available. A new way of learning is therefore needed to allow for efficient model fitting in the 21st century. The birth of compressive learning in recent years has provided a novel solution to the bottleneck of learning from big data. Situated at the core of the compressive learning framework is the construction of a so-called sketch. The sketch is a compact representation of the data that provides sufficient information for specific learning tasks. In this thesis we develop the compressive learning framework to a host of new models and applications. In the first part of the thesis, we consider the group of semi-parametric models and demonstrate the unique advantages and challenges associated with creating a compressive learning paradigm for these particular models. Concentrating on the independent component analysis model, we develop a framework of algorithms and theory enabling magnitudes of compression with respect to memory complexity compared to existing methods. In the second part of the thesis, we develop a compressive learning framework to the emerging technology of single-photon counting lidar. We demonstrate that forming a sketch of the time-of-flight data circumvents the inherent data-transfer bottleneck of existing lidar techniques. Finally, we extend the compressive lidar technology by developing both an efficient sketch-based detection algorithm that can detect the presence of a surface solely from the sketch and a sketched plug and play framework that can integrate existing powerful denoisers that are robust to noisy lidar scenes with low photon counts.

Lay Summary

The availability of data has increased exponentially in recent years, partly thanks to the advances in technology and the growth of data centric industries. However, many of the models used to uncover the relationships in the data require memory and computational resources that typically scale with the size of the datasets. As a result, training models on a large scale has become increasingly slow and expensive. Recently, the field of compressive learning has attempted to tackle this problem by compressing the data into a fixed sized summary, called a sketch. Crucially, the sketch retains sufficient salient information of the data enabling a specific model to be trained directly on the compact sized sketch without recourse to the original dataset. In this thesis, we extend the compressive learning paradigm by developing new models and applications. In Part I, we develop a compressive independent component analysis scheme that enables one to find mutually independent components of the dataset using only a sketch resulting in significant memory compression. Part II of the thesis concentrates on the application of light detection and ranging (lidar), a 3D imaging modality that achieves high resolution depth images using eye-safe lasers. By developing a sketch based lidar framework, it is shown that the data associated with lidar scenes can be enormously compressed without sacrificing the overall resolution of the depth images. This paves the way for high resolution, high frame-rate lidar devices that can be processed in real time.

Declaration of originality

I hereby declare that the research recorded in this thesis and the thesis itself was composed and originated entirely by myself in the Institute of Digital Communication (IDCOM) at The University of Edinburgh.

Michael Patrick Sheehan

Acknowledgements

I thank my supervisor Prof. Mike Davies for his support throughout my studies as well as his ongoing enthusiasm and energy towards research that is inspiring. I thank him for the opportunity, freedom and encouragement that he has provided. It has been a privilege to spend the last few years learning a great deal from him.

My colleague and friend, Dr. Julian Tachella, has been instrumental in this project. He has been a great source of knowledge and provided a fresh perspective on research helping me to learn along the way. Dr. Madeleine Kotzagiannidis contributed towards the early stages of the project and I thank her for the constructive discussions and the mentoring she has provided.

I also give thanks to my examiners Prof. Laurent Jacques and Dr. Nick Polydorides for their time evaluating the manuscript, and their useful feedback and critique, which has allowed me to significantly improve the quality of this thesis.

I have some incredible support throughout the project. I thank Tessa for her unrelenting love and encouragement. To my parents, thank you for everything. I am forever indebted to you for giving me the opportunities and experiences that have made me who I am. Finally to my whole family, I couldn't ask for a better bunch.

Contents

Lay Summary . . . iii
Declaration of originality . . . iv
Acknowledgements . . . v
Contents . . . vi
List of figures . . . ix
List of tables . . . xiii
Acronyms and abbreviations . . . xiv
Nomenclature . . . xv

1 Introduction **1**
1.1 Motivation . . . 1
1.1.1 Large Scale Learning Challenges . . . 2
1.1.2 Compressive Learning to the Rescue? . . . 3
1.1.3 Contributions . . . 4
1.1.4 List of Publications . . . 7

2 Background **9**
2.1 Parameter Inference . . . 10
2.1.1 Probability Density Fitting . . . 12
2.1.2 Maximum Likelihood Estimation . . . 13
2.1.3 Generalized Method of Moments . . . 14
2.1.4 Semi-Parametric Learning . . . 15
2.1.5 Principal Component Analysis . . . 16
2.1.6 Independent Component Analysis . . . 17
2.2 Estimation Theory . . . 20
2.2.1 Bias and Consistency . . . 21
2.2.2 Central Limit Theorem . . . 21
2.2.3 Efficiency . . . 22
2.2.4 Sufficiency . . . 23
2.3 Compressive Learning . . . 25
2.3.1 Compressive Sensing . . . 25
2.3.2 Principles of Compressive Learning . . . 32
2.3.3 Existing Compressive Learning Models . . . 38
2.3.4 Advantages and Disadvantages of CL . . . 40
2.4 Large Scale Learning . . . 41
2.4.1 Sub-Sampling . . . 42
2.4.2 Dimensionality Reduction . . . 45
2.4.3 Sketching . . . 46
2.5 Single Photon Counting Lidar . . . 48
2.5.1 Data Transfer Bottleneck . . . 51

I Part One: New Models **54**

3 Compressive Independent Component Analysis **55**
3.1 Introduction . . . 55
3.2 Compressive Learning Principles for Cumulant ICA . . . 57
3.3 Compressive Independent Component Analysis Theory . . . 60
3.3.1 Proof of Theorem 1 . . . 62
3.3.2 Finite Sample Effects . . . 63
3.3.3 Discussion on Unwhitened Data . . . 64
3.4 CICA Algorithms . . . 65
3.4.1 Iterative Projection Gradient . . . 65
3.4.2 Unwhitened IPG . . . 67
3.4.3 Alternating Steepest Descent . . . 68
3.5 Empirical Results . . . 71
3.5.1 Phase Transition . . . 71
3.5.2 Statistical Efficiency . . . 72
3.5.3 Cylinder Velocity Field . . . 75
3.6 Concluding Remarks . . . 76
3.A Appendix . . . 78
3.A.1 Proof of Lemma 2 . . . 78
3.A.2 Proof of Theorem 2 . . . 89
3.A.3 Conic Properties . . . 90
3.A.4 1st Order Differential . . . 91
3.A.5 Second Order Differential . . . 91

4 Compressive Learning for Semi-Parametric Models **93**
4.1 Introduction . . . 93
4.2 Motivation . . . 94
4.2.1 Topology of Semi-Parametric Models . . . 96
4.3 Compressive Learning Framework for Semi-Parametric Models . . . 97
4.4 When Does Compressive Semi-Parametric Learning Work? - A Case Study . . . 99
4.4.1 Compressive Generalized Principal Component Analysis . . . 100
4.4.2 Comparison . . . 103
4.5 Discussion . . . 104
4.6 Concluding Remarks . . . 106

II Part Two: Applications **107**

5 A Sketching Framework for Single Photon Counting Lidar **108**
5.1 Introduction . . . 108
5.2 Sketched Lidar . . . 111
5.2.1 Compressing Single Depth Data . . . 111
5.2.2 Sampling the ECF . . . 114
5.2.3 Practical Hardware Considerations . . . 115
5.3 Sketched Lidar Reconstruction . . . 116
5.3.1 Statistical Estimation . . . 116

5.3.2 Central Limit Theorem . . . 118
5.3.3 Statistical Efficiency . . . 119
5.4 Experiments . . . 126
5.4.1 Experimental set up . . . 126
5.4.2 Synthetic Data . . . 127
5.4.3 Real Data . . . 129
5.4.4 Wordlength Considerations . . . 136
5.5 Concluding Remarks . . . 138
5.A Deriving the Circular Mean Estimate From the ECF Estimation . . . 138
5.B Photon Starved Regime . . . 139
5.C Comparison to the iFFT approach . . . 140

6 Robust Detection and Real-Time Processing of Sketched Single-Photon Counting Lidar 143
6.1 Introduction . . . 143
6.2 Pixelwise Surface Detection . . . 145
6.2.1 Sketch-Based Detection Algorithm . . . 147
6.2.2 Spatial Regularization . . . 148
6.2.3 Empirical Results . . . 149
6.2.4 Discussion . . . 151
6.3 Multipixel Sketched Lidar . . . 153
6.3.1 Experiments . . . 154
6.4 Concluding Remarks . . . 157

7 Conclusion and Future Perspectives 159
7.1 Conclusions . . . 159
7.2 Future Perspectives . . . 161
7.2.1 Efficient One-Stage Compressive ICA . . . 161
7.2.2 Alternative Compressive Semi-Parametric Learning . . . 161
7.2.3 Compressive Transfer Learning . . . 162
7.2.4 On-Chip Implementation of Sketched Lidar . . . 162
7.2.5 Complex Lidar Scenarios . . . 163
7.2.6 Extensions to Other Photon Counting Imaging Modalities . . . 164

References 166

List of figures

1.1 A schematic illustrating a simplified version of the machine learning process. . 2
1.2 A schematic illustrating the compressive learning process. 4

2.1 An asymptotically efficient estimator achieves the CRLB as N becomes large. . 23
2.2 Schematic of the restricted isometry property (left) and the instance optimal decoder property (right). 27
2.3 ℓ_p norms for $p = 1, 2$ in 2D. The red line represents the linear constraints induced by the observations. A solution of the ℓ_p problem occurs when the linear constraints intersects with the ℓ_p ball. The ℓ_1 problem induces a sparse solution i.e. the intersection occurs on the axis. 28
2.4 The core principles of compressive learning split into the sketching phase and the learning phase. 32
2.5 The three main approaches used to tackle the issues surrounding large scale learning. 42
2.6 Single photon lidar: a laser is directed at a pixel in a scene which consists of a semi-transparent camouflage netting with a human stood behind. The recorded photons produce a TCSPC histogram that exhibits 2 peaks of varying intensity for the camouflage surface (1) and human (2), as well as spurious photon detections from ambient sources like the sun (3). 49
2.7 A schematic of a typical lidar pixel where either one or multiple SPADs and TDCs are used. 50

3.1 A phase transition between unsuccessful and successful mixing matrix inference as the sketch size m increases and the number of independent components is fixed at $n = 8$. 72
3.2 A phase transition between unsuccessful and successful mixing matrix inference as the sketch size m and the number of independent components n increases. 73
3.3 (a) Student's t distribution $(\nu = 3)$ (b) Laplace distribution $(\mu = 0, b = 1)$ (c) continuous uniform distribution $(a = -\sqrt{3}, b = \sqrt{3})$ (d) mixture of 2 Laplaces $(\mu_1, \mu_2 = -1, 1\ b_1 = b_2 = 1)$ (e) symmetric bimodal mixture of Gaussians $(\mu_1, \mu_2 = -1, 1\ \sigma_1 = \sigma_2 = 0.15)$ (f) asymmetric unimodal mixture of Gaussians $(\mu_1, \mu_2 = -0.7, 0.5\ \sigma_1 = \sigma_2 = 0.5)$ 74
3.4 The relative efficiency of the full data cumulant tensor (Comon's ICA) and sketch mixing matrix estimates for increasing sketch size m. 75
3.5 The figure shows the velocity field around a cylinder for a fixed point in time. . 76
3.6 From left to right the dominant fluctuations of the streamwise velocity field. From top to bottom the Fast ICA, JADE, Comon and CICA reconstructions. . . 77
3.7 The figure shows the effect of the sketch size on the reconstruction of the fluctuations. From top to bottom a sketch size of $m = 144, 108$ and 72. 77

4.1 A schematic diagram detailing the original compressive learning framework proposed in [1]. 95

4.2 A schematic diagram of the probability equivalence class whereby many distributions collapse down to one point on the intermediary set of statistics. 97
4.3 A schematic diagram detailing the reformulated compressive learning framework for semi-parametric models. 99
4.4 A set of data points in $\mathbb{R}^3$ drawn from a mixture of three distributions that supported on the plane P_1 ($d_1 = 2$), the line L_1 ($d_2 = 1$) and the line L_2 ($d_3 = 1$). 101
4.5 The compression ratio $\frac{DR}{Nd}$ for compressive GPCA for a rank of $R = 0.05D$ (top) and $R = 0.8D$ (bottom). 104
4.6 A schematic diagram of parametric compressive learning (left) and semi-parametric compressive learning (right). 106

5.1 An original zoomed in (around the target) TCSPC histogram with $T = 4613$ bins (left) and the coarse version of 50 bins (right) 109
5.2 The ground truth 3D depth image of a polystyrene head (left) and reconstruction using 50 coarse bins (right). The coarse binning method suffers from the staircase effect. 110
5.3 The TCSPC histogram with $t_1 = 320$. The circular mean estimate (yellow) and the standard mean estimate (red) superimposed. 113
5.4 Histograms of the estimation error $(\hat{t} - t_1)$ for increasing photon count N where the sketched lidar estimate (circular mean) is denoted by $\hat{t}$. The expected error distribution in Eqn 5.9 is depicted in red. 120
5.5 The CF (top) of a short (blue solid) and long (red dashed) tailed impulse response function (bottom). 122
5.6 The REP as a function of the number of real measurements ($2m$) for a single peak lidar scene. 123
5.7 The REP as a function of the number of real measurements ($2m$) for a lidar scene with 2 surfaces. 124
5.8 Comparison of the RMSE achieved by wide and narrow Gaussian pulse width coarse binning to our proposed SMLE algorithm. 125
5.9 RMSE level set contour plots for varying SBR levels and number of real measurements $2m$ for a photon count of $N = 100$. The RMSE level are $10\Delta\tau$ (left) and $2\Delta\tau$ (right). The legend is defined for both plots. 128
5.10 RMSE level set contour plots for varying SBR levels and number of real measurements $2m$ for a photon count of $N = 1000$. The RMSE level are $10\Delta\tau$ (left) and $2\Delta\tau$ (right). The legend is defined for both plots. 129
5.11 RMSE level set contour plots for varying SBR levels and number of real measurements $2m$ for a photon count of $N = 100$ for detecting 95% of peaks within the level sets of $10\Delta\tau$ (left) and $3\Delta\tau$ (right). The legend is defined for both plots. 130
5.12 RMSE level set contour plots for varying SBR levels and number of real measurements $2m$ for a photon count of $N = 100$ for detecting 95% of peaks within the level sets of $10\Delta\tau$ (left) and $3\Delta\tau$ (right). The legend is defined for both plots. 131
5.13 The CF (bottom) of the data driven impulse response function (top) of the polystyrene head dataset. 131

5.15 The RMSE as a function of the number of real measurements ($2m$) for the polystyrene head dataset. 132
5.16 The CF (bottom) of the data driven impulse response function (top) of the camouflage dataset. 133
5.14 The face dataset lidar reconstructions of the sketched lidar and coarse binning method for the real valued measurement size $2, 8, 20$. Both the cross correlation (XCORR) reconstruction and the ground truth image are given for comparison. 134
5.17 The camouflage dataset lidar reconstructions of the sketched lidar and coarse binning method for the real valued measurement size ($2m$) of $2, 8, 20$. Both the cross correlation (XCORR) reconstruction and the ground truth image are given for comparison. 135
5.18 The reconstructions of the polystyrene head dataset for a sketch of size $m = 2, 5$ and 10 where each sketch entry has a precision of 32, 16, 12, 8 and 4 bits. 136
5.19 Mean absolute error of the reconstructions for a sketch of size $m = 2, 5$ and 10 for individual sketch entries of 4,8,12,16 and 32 bit precision. 137
5.20 Mean absolute error of the reconstructions for a sketch of size $m = 2, 5$ and 10 for the total wordlength of the whole sketch. 137
5.21 Sketched Lidar performs comparatively well (in terms of RMSE) compared with the full data approaches of cross correlation (XCORR) and maximum peak estimation in the photon starved regime. 140
5.22 Comparison of the depth reconstruction of sketched lidar and cross correlation (XCORR) using the RMSE ratio R for varying SBR levels and photon counts in the photon starved regime. Sketched Lidar performs favourably compared to XCORR for the majority of SBR values. 141
5.23 Comparison of the depth reconstruction of sketched lidar and the iFFT method using the RMSE ratio R for varying SBR levels and photon counts for a non-Gaussian asymmetric IRF. Sketched Lidar performs equally or favourably to iFFT for all SBR and photon count pairs. 142

6.1 A TCSPC histogram of a pixel containing no informative surface peak. 145
6.2 A 3D image of a face (left) with its associated pixelwise surface detection map (right). 145
6.3 Detection performance of the sketch-based method for sketch sizes of $m = 3, 5, 10$, the coarse histogram test for histograms of size $T_r = 10, 50, 100$, and the full data K-S test. The graphs correspond to a detection probability of 95%. 150
6.4 Empirical probability of detection for the proposed sketch-based detection scheme using a sketch size $m = 10$. 151
6.5 Probability of false alarm of the sketch-based method for sketch sizes of $m = 3, 5, 10$, the coarse histogram test for histograms of size $T_r = 10, 50, 100$, and the full data K-S test. 152
6.6 Detection maps of the polystyrene dataset [2] for the proposed sketch and sketch plus TV detection schemes in comparison with other non-compression detection techniques. 152
6.7 Empirical probabilities of detection (top) and false alarm (bottom) for the evaluated detection methods using the polystyrene head dataset. 153

6.8 Execution time of RT3D [3] and the proposed sketched SRT3D as a function of the mean number of photons per pixel. RT3D suffers from a linear complexity, whereas SRT3D only depends on the size of the sketch m. 155
6.9 3D reconstructions obtained by the proposed sketched RT3D algorithm for different sketch sizes m and other competing algorithms. The proposed SRT3D method incorporates spatial regularization, providing stable reconstructions in settings with low SBR or low number of measured photons. 156
6.10 Performance of the evaluated algorithms for the polystyrene head dataset with SBR=1. 156
6.11 Reconstruction of the scene in [3] with 2 surfaces per pixel by the original RT3D and its sketched version. Using a sketch size of only $m = 10$ is enough to provide the same reconstruction quality. 157

List of tables

4.1 A summary of the current semi-parametric compressive learning models and their properties. 105

6.1 Probabilities of detection (PD) and probabilities of false alarm (PFA) for the proposed sketch-based detection schemes and other detection algorithms. The sketch size is set at $m = 5$ and the full data χ^2 test was chosen using 50 adjacent bins to optimise the PD/PFA trade-off. 151

6.2 Execution time in milliseconds for different scene sizes in pixels obtained by the proposed sketched RT3D algorithm for a sketch size of $m = 5$ and $m = 10$, respectively. 157

Acronyms and abbreviations

ASD	Alternating Steepest Descent
CF	Characteristic Function
CICA	Compressive Independent Component Analysis
CL	Compressive Learning
CLT	Central Limit Theorem
CS	Compressive Sensing
CUE	Continuous Updating Estimator
ECF	Empirical Characteristic Function
FFT	Fast Fourier Transform
ICA	Independent Component Analysis
IPG	Iterative Projected Gradient
IRF	Impulse Response Function
GeMM	Generalized Method of Moments
GMM	Gaussian Mixture Model
GPCA	Generalized Principal Component Analysis
Lidar	Light Detection and Ranging
MLE	Maximum Likelihood Estimator
PCA	Principal Component Analysis
PD	Probability of Detection
PFA	Probability of False Alarm
RIP	Restricted Isometry Property
RMSE	Root Mean Squared Error
SBR	Signal to Background Ratio
SRHT	Subsampled Randomized Hadamard Transform
SPAD	Single Photon Avalanche Diode
TCSPC	Time Correlated Single Photon Counting
TDC	Time-to-Digital Converter
ToF	Time of Flight
TV	Total Variation

Nomenclature

$\langle \cdot, \cdot \rangle$	Inner product
$(\cdot)^T$	Transpose
$\times_j$	Tensor - matrix multiplication along the jth mode
$\otimes$	Kronecker product
$\mathbb{E}(\cdot)$	Expectation
$\mathbb{R}^d$	d-dimensional Euclidean space
$\|\cdot\|_0$	ℓ_0 "norm"
$\|\cdot\|_1$	ℓ_1-norm
$\|\cdot\|_2$	ℓ_2-norm
$\|\cdot\|_{\text{TV}}$	Total variation norm
$\|\cdot\|_{1,2}$	$\ell_{1,2}$-norm
$\|\cdot\|_*$	Nuclear norm
$\mathbb{S}^{d-1}$	Unit sphere in $\mathbb{R}^d$ centered at the origin
$\mathfrak{S}_\theta$	Model set parameterized by θ
$\mathfrak{C}$	Space of 4th order cumulant tensors
$\mathcal{N}(\cdot, \cdot)$	Normal distribution
$\nabla f(\cdot)$	Gradient operator of f
$\partial f(\cdot)$	Sub-differential operator of f
$\mathfrak{R}$	Regularization function
λ	Regularization parameter
$\mathcal{A}$	Sketching operator
$\text{vec}(\cdot)$	Vectorization operator
$\Psi_{\mathcal{P}}$	Characteristic function of the distribution $\mathcal{P}$
Δ	Decoder
$\mathcal{P}_{\mathcal{K}}(\cdot)$	Orthogonal projection towards a constraint set $\mathcal{K}$

Chapter 1
Introduction

Data is the new oil

Clive Robert Humby

1.1 Motivation

In the early 2000s, famous British mathematician and data scientist Clive Humby claimed that data is the new oil. Fast forward twenty years to the midst of the so-called *data revolution* and his quote is as pertinent as ever. With data comes the potential to uncover patterns and enable powerful predictions. From sports and marketing through to science and finance, most domains of industry are utilising this widely available commodity to gain invaluable insights. In recent years the volume, frequency and availability of data has rapidly increased with little sign of slowing. Many experts have coined this exceptional growth as the *big data* era. So what has fuelled the big data era? The main two drivers can be attributed to the invention of the internet and major advances in technology. The success of the internet has allowed humankind to become more connected and live more conveniently due to the prominence of social media, search engines and online shopping. With each online transaction, a digital footprint is harvested and recorded as data. On the other hand, the exponential growth of technology has resulted in products such as mobile phones, autonomous vehicles and personal computers consisting of multiple sensors which can record data at ever faster rates. In addition, such technology has become increasingly affordable and accessible in recent years. All of these factors have lead to ever increasingly large datasets.

Synonymous with big data and shadowing its rise is the field of machine learning which concerns the problem of learning the underlying process of a task, such as classification, detection or prediction, given a set of data. Here machine learning refers to the umbrella terminology that encompasses any *model* that is trained using a collection of data, including the broader family of deep learning. In its most simplified form, machine learning consists of a model f_θ

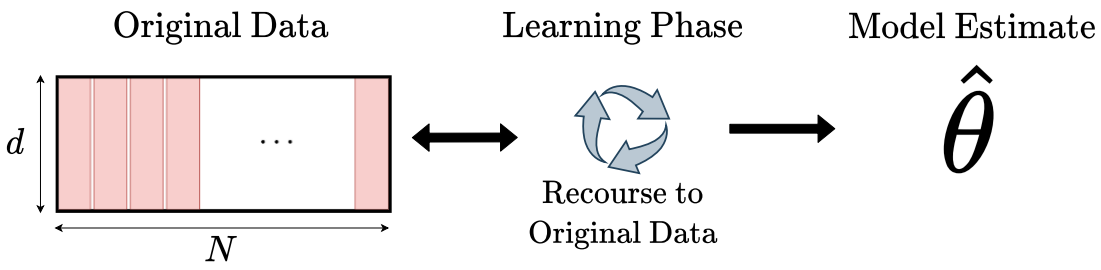

Figure 1.1: *A schematic illustrating a simplified version of the machine learning process.*

parametrized by θ, whereby a collection of data $\mathcal{X}$ is used to learn an estimate $\hat{\theta}$ of the model. The learning stage consists of optimisation algorithms that typically require recourse to the original dataset. Figure 1.1 depicts a simplified schematic of the machine learning process.

Both the rise of big data and machine learning are intertwined. Conventional statistical wisdom states that larger datasets enable practitioners to build increasingly accurate models. As a result, machine learning facilitates a greater demand for big data. The self-perpetuating growth of both big data and machine learning poses significant challenges for modern day large scale learning.

1.1.1 Large Scale Learning Challenges

The challenges of large scale learning can be characterised into two main groups, namely resource limitations and societal costs.

Resource Limitations

Datasets typically assume the form of a matrix (once pre-processed) as shown in Figure 1.1, consisting of N individual entries (representing for example a patient, product or specific point in a time-series) with a dimension or feature space d containing specific attributes of the entry (for example height, cost, temperature). With ever increasing technology, memory access and sensor design the size of both N and d have increased substantially over the years. It is commonplace to have datasets that consist of millions of entries N with thousands of associated features d, requiring up to terabytes of available storage. However, many of the classical optimisation algorithms associated with learning the model f_θ require recourse to the original dataset which poses a restriction as the dataset must typically be stored on local memory. Secondly, the computational complexity, or the runtime of learning the model f_θ generally scales with the size of N and d, regularly exhibiting a polynomial dependency on these dimensions. Large scale learning can therefore become extremely time intensive, taking days or even weeks

to complete. For this reason, large scale learning quickly becomes prohibitive.

Societal Costs

Associated with the theoretical resource limitations are the increasing costs to society. Due to the large computational overhead of large scale learning, practitioners and industry frequently resort to graphic processing units (GPUs) that parallelize numerical operations and can substantially reduce the runtime of their optimisation algorithms. However, these units consume enormous amounts of energy and are therefore expensive to run from a monetary point of view. Furthermore, they significantly contribute to CO_2 emissions, for instance a cluster of eight RTX2080Ti NVIDIA GPUs used over a 24 hours duration emits approximately 35.5 kg of CO_2 emissions which is equivalent to driving an average car for a total of 143.2 kilometres [4]. Due to the increased demand for quicker computations, many GPUs are updated within a short time frame producing a greater quantity of waste and leading to inflated prices due to GPU shortages. Owing to the increased monetary cost and the restrictive access to big datasets, large scale learning has become progressively limited to the larger corporations. For example, *OpenAI* trained GPT-3, a state-of-the-art language model consisting of 125 billion trainable parameters, at a reported cost of $\$12$ million [5]. As a result, this enables the larger corporations to hold a near monopoly as smaller institutions, including university research groups, can not feasibly compete.

Traditional approaches to large scale learning are not sustainable and therefore a new learning paradigm is needed. Does *compressive learning* provide the answer?

1.1.2 Compressive Learning to the Rescue?

Compressive learning (or often referred to as sketched learning) partially addresses the fundamental challenges of large scale learning by severely compressing the whole dataset into a compact representation of fixed size m, named a so-called sketch, in a single (or limited) pass of the data prior to learning. In general, the sketch is constructed by taking an empirical average of a nonlinear function over the dataset. Once the sketch is formed, the parameters of the model are inferred solely from the sketch, hence a compressive learning algorithm, for a given task or model, needs never to return to the original dataset. The sketching and learning stages of compressive learning are illustrated by the schematic in Figure 1.2. Fundamentally, the size of the sketch does not scale with the dimensions of the dataset, or indeed the data's underlying

dimensionality, but instead is driven by the complexity or dimensionality of the learning task or model of interest. In theory, one can work with datasets of arbitrary length, as the dimension of the sketch is fixed constant throughout, making compressive learning especially amenable to large scale learning. The key advantages of compressive learning are

- Only the sketch of size $m \ll Nd$ is required throughout the learning stage of compressive learning, therefore recourse to the original dataset is not required and subsequently can be discarded from local memory.
- As the sketch is constructed by computing the empirical mean of a nonlinear function, compressive learning is highly amenable to both parallel processing and distributed learning. In addition, due to its nature, the sketch can be computed on the fly without any additional computational overhead (as will be shown in Chapter 5).
- The compressive learning algorithms that estimate the parameters θ of the model directly from the sketch typically have a computational complexity that is dependent only on the size of the sketch m, thereby significantly reducing the computational burden of learning.

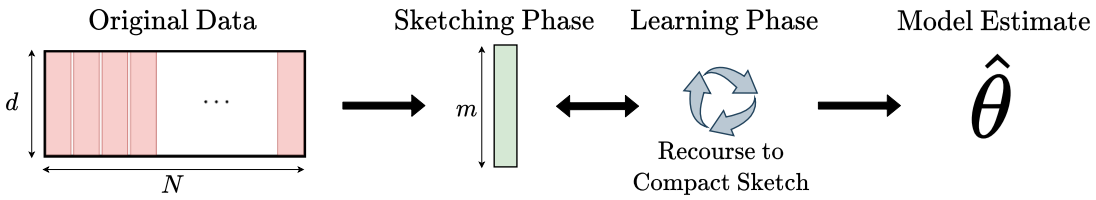

Figure 1.2: *A schematic illustrating the compressive learning process.*

In this thesis, we develop the compressive learning framework by introducing new models and applications which provide a solution to some of the challenges faced by modern day large scale learning. The contributions of the thesis are stated below.

1.1.3 Contributions

This thesis is divided into 2 parts, namely, new compressive learning models and compressive learning applications. To date, compressive learning has mainly centred around building frameworks for parametric models including the Gaussian mixture and K-means models. In this thesis, we develop the compressive learning framework to explicitly highlight the intricacies and challenges that are unique to the class of semi-parametric models. The key novelty in this first

half of the thesis is the introduction of a compressive independent component analysis (ICA) framework in Chapter 3, which includes both theory and practical algorithms allowing for substantial space complexity compression compared with existing ICA methods. In Chapter 4, we reformulate the compressive learning framework in [1, 6] to be explicit in the construction of a compressive semi-parametric scheme. With the use of a case study between the models of ICA and generalized principal component analysis (GPCA), the unique challenges of compressive semi-parametric learning are highlighted and we demonstrate when efficient compression can be attained. Our reformulation provides a clear blueprint on constructing compressive learning schemes for other semi-parametric models.

The second half of the thesis focuses on the application perspective of compressive learning. The key novelty of this part is the development of a compressive time-of-flight (ToF) imaging framework which utilises a compressive mixture model in its construction. In Chapter 5, we provide a framework in which a sketch of the ToF data allows efficient reconstruction of a field of view depth scene. In Chapter 6, we further develop the framework by designing a pixel-wise object detection scheme that requires only a small sized sketch and also a multi-pixel denoiser that is robust to noisy scenes. This thesis contains a background chapter (Chapter 2), 4 technical chapters (Chapters 3-6) and a conclusion chapter (Chapter 7). A brief overview of each technical chapter is highlighted below.

Chapter 3: Compressive Independent Component Analysis

In this chapter we build a compressive ICA framework including both theory and algorithms. Focusing on cumulant based ICA, we show that solutions to the cumulant based ICA model have particular structure that induce a low dimensional model set that resides in the larger cumulant tensor space. One of the main contributions of this chapter is proving that a restricted isometry property (RIP) holds for the sketch of cumulant tensors constructed using Gaussian ensembles. The established RIP permits a robust instance optimal decoder and therefore a tractable compressive ICA scheme. The other main contribution of this chapter is the proposal of two distinctively different algorithms in the form of an alternating steepest descent (ASD) and iterative projected gradient (IPG) method which are shown to be robust and efficient at estimating the parameters of the ICA model. Through extensive simulations on both synthetic and real data, we empirically show the specific phase transition between unsuccessful and successful ICA estimation as a function of the sketch size and demonstrate that the order of compression asserted from the RIP theory is realised through the empirical results.

Chapter 4: Compressive Learning for Semi-Parametric Models

In this chapter, we reformulate the original compressive learning framework in [1, 6] to explicitly cater for the unique class of semi-parametric models. One of the unique characteristics of semi-parametric models is that they are not fully parametrized. As a result, we leverage some intermediate space of statistics that allow identifiability of the semi-parametric model. However, the dimensionality of the finite intermediate space can result in a compressive learning scheme that quickly becomes infeasible due to the required sketch size. The main contribution of this chapter is the reformulation of the existing compressive learning framework and to provide a case study between the models of ICA and GPCA, demonstrating when compressive semi-parametric can fail and succeed.

Chapter 5: A Sketching Framework for Single Photon Counting Lidar

Single photon counting light detection and ranging (lidar) is a prominent depth imaging tool used extensively in the automobile, defense and agriculture industries. A major data transfer bottleneck arises on the lidar device when either the number of photons detected per pixel is large or the time-stamp resolution is fine, which is becoming more apparent due to the major advances in hardware capabilities. The major contribution of this chapter is the development of a robust compressive time-of-flight framework that circumvents the data transfer bottleneck of modern lidar devices whilst achieving efficient and accurate depth estimation. A mixture model consisting of both signal and background photon components is utilised in the framework. It is demonstrated on both synthetic and real datasets that the size of the sketch needs only to be of the order of the number of surfaces in the scene. In most cases, there is only 1 surface and therefore the compression realised in comparison to existing techniques is substantial. Another main contribution of this chapter explores the loss of information incurred by taking a sketch of a certain size. It is shown that even at low signal-to-background ratios (SBR), only a modest sketch size is required to incur negligible information loss.

Chapter 6: Robust Detection and Real-Time Processing of Sketched Single-Photon Counting Lidar

Detecting the presence of an object or surface for each pixel in a scene is important for many downstream tasks and can also further reduce the computational and memory load by omitting non-informative peaks (e.g. pixels with no surface present) from further processing. However, existing techniques typically scale at best $\mathcal{O}(T \log T)$, where T is the temporal resolution of the

system. This can become extremely slow for devices that permit high temporal resolution. The first main contribution of this chapter is to extend the compressive lidar framework to enable accurate detection based solely on the compact representation sketch. It is demonstrated on both synthetic and real data that the accuracy, measured in terms of probability of detection (PD) and probability of false alarm (PFA), is competitive with the state-of-art, while exhibiting a computation complexity of $\mathcal{O}(m)$, which is at a fraction of the cost of the state-of-the-art.

Our second main contribution of this chapter is to extend the initial sketched lidar reconstruction algorithm proposed in Chapter 5 by designing a plug and play multi-pixel denoiser algorithm that exploits the spatial correlation of neighbouring pixels to improve the reconstruction performance. It is demonstrated on real and synthetic data that the proposed sketched multi-pixel denoiser is robust to both low photon counts and low SBR scenes. Importantly, it is highlighted that one can easily develop sketched versions of other existing and future lidar denoisers by employing a plug and play format of replacing the standard data fidelity term by a specific sketch based cost function.

1.1.4 List of Publications

This thesis is based on the following peer-reviewed publications, preprints and patents during my PhD studies:

Chapter 3

- **Michael P. Sheehan**, Madeleine S. Kotzagiannidis, and Mike E. Davies. "Compressive Independent Component Analysis." 2019 27th European Signal Processing Conference (EUSIPCO). IEEE, 2019.

- **Michael P. Sheehan**, and Mike E. Davies."Compressive Independent Analysis: Theory and Algorithms." Information and Inference: A Journal of the IMA, 2022

Chapter 4

- **Michael P. Sheehan***, Antoine Gonon*, and Mike E. Davies."Compressive Learning for Semi-Parametric Models." 2019, arXiv preprint arXiv:1910.10024 (2019). * The authors have contributed equally to the paper.

Chapter 5

- **Michael P. Sheehan**, Julián Tachella, and Mike E. Davies."A Sketching Framework for Reduced Data Transfer in Photon Counting Lidar." 2021 Transactions on Computational Imaging, IEEE 2021.
- Mike E. Davies, **Michael P. Sheehan**, Julián Tachella, (2022) Sketching Technique for Reduced Data Transfer, International Patent Application No. PCT/GB2022/050263

Chapter 6

- **Michael P. Sheehan**, Julián Tachella, and Mike E. Davies. "Surface Detection for Sketched Single Photon Lidar." 2021 29th European Signal Processing Conference (EUSIPCO). IEEE, 2021.
- Julián Tachella, **Michael P. Sheehan**, and Mike E. Davies. "Sketched RT3D: How to reconstruct billions of photons per second." International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE 2022.

Chapter 2
Background

In this chapter, we will present an overview on parameter inference and compressive learning that we will build upon in the four technical chapters of this thesis. This will begin in Section 2.1 where a description of various forms of parameter estimation are provided, including probability density estimation and distribution-free learning. The role of empirical characteristic function estimation as a parameter inference tool will be detailed in Section 2.1.3, which will be later used in Chapter 5 and 6 to build a compact representation of the lidar data. The class of semi-parametric models will be detailed in Section 2.1.4 where we exploit their unique structure to develop a compressive learning framework for semi-parametric models in Chapter 4. In Section 2.1.6 we present the learning task of independent component analysis and demonstrate that it belongs to the class of semi-parametric models, which we will use in Chapter 3 to construct a compressive independent component analysis scheme that includes both practical algorithms and theoretical results.

Next, we present some key properties of estimators in Section 2.2, that will be used to design practical algorithms and analyse the proposed methods in this thesis. The notion of a bias estimator will be described in Section 2.2.1, which will be used to design a sampling scheme in Chapter 5 that produces unbiased estimates of the lidar parameters. The role of efficiency will be detailed in Section 2.2.3, which we will later use in both Chapter 3 and 5 to analyse the loss of information incurred by computing a compact representation of the data. The central limit theorem will also be introduced in Section 2.2.2, which will be leveraged extensively in Part II to design both a estimation and detection algorithm for single photon lidar.

The fundamentals of compressive learning will be provided in Section 2.3 that we build upon throughout the technical chapters. Section 2.3.1 provides a review of the well-established field of compressive sensing which provides inspiration to much of the principles of compressive learning that will be introduced in Section 2.3.2. In addition, many of the theoretical tools and algorithms that will be presented in Section 2.3.1 will be utilised in Chapters 3 and 4 due to the inherent structure of semi-parametric models. A review of existing compressive

learning methods is presented in Section 2.3.3 and the main advantages and disadvantages of compressive learning are discussed in Section 2.3.4.

A review on other methods that tackle the complexities associated with large scale learning is presented in Section 2.4. The techniques of sub-sampling, dimensionality reduction and linear sketches are introduced and compared to the compressive learning method, where we highlight their main limitations.

Finally, we provide a description of single photon counting lidar in Section 2.5 and detail the data transfer bottleneck inherent in many modern day lidar devices in Section 2.5.1. In Chapter 5, we will develop a sketching framework for reduced data transfer that reduces the computational and data transfer complexities associated with high resolution, high frame rate lidar devices without sacrificing the quality of the reconstructed images.

2.1 Parameter Inference

The goal of compressive learning and therefore this thesis is to provide a solution to large-scale inference problems in an efficient manner that reduces computational and memory complexities. In Chapter 1, the notion of a learning model f that attempts to describe a specific learning task from a given set of observations was broadly introduced. In this section, we make these definitions explicit and introduce the specific class of learning models we will be considering throughout the later chapters of this thesis.

There are 3 main components to any learning task, namely, data, a learning machine and a measure of discrepancy. Each component, as defined below, will allow us to define the goal of statistical learning.

Data

To learn a model for a given task, one must have realisations from the true data generating distribution. Let $\mathbf{X}$ denote the finite set of data points

$$\mathbf{X} \coloneqq \{\mathbf{x}_i \in \mathcal{X} \mid i = 1, \dots, N\}, \tag{2.1}$$

where the data belongs to some signal space $\mathcal{X} \subseteq \mathbb{R}^d$. As discussed in Chapter 1, a dataset is typically concatenated into a matrix $\mathbf{X} \in \mathbb{R}^{d \times N}$ where the number of columns N and the rows d define the length and dimensionality of the dataset, respectively. In statistical learning theory [7] it is assumed that the samples $\mathbf{x}_i$ are independent and identically distributed (i.i.d) realisations of some unknown data generating distribution $\mathcal{P}_0 \in \mathcal{P}(\mathcal{X})$:

$$\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N \overset{\text{i.i.d}}{\sim} \mathcal{P}_0, \tag{2.2}$$

where $\mathcal{P}(\mathcal{X})$ denotes all possible probability distributions over $\mathcal{X}$.

Learning Machine

The next main component to a learning task is the choice of learning machine. A learning machine implements a class of models $f(\mathbf{x}, \theta)$ parameterized by a set of parameters $\theta \in \Theta$ where $\Theta \subseteq \mathbb{R}^p$ denotes the parameter set [7]. For shorthand, we denote $f_\theta(\mathbf{x}) = f(\mathbf{x}, \theta)$. Each model defines the map

$$f_\theta : \mathcal{X} \mapsto \mathcal{Y} \tag{2.3}$$

to an output space $\mathcal{Y}$. Depending on the learning task, the space $\mathcal{Y}$ could be, for example, $\mathcal{Y} = \mathbb{R}^k$ (regression), $\mathcal{Y} = \{1, 2, \ldots, C\}$ (classification) or $\mathcal{Y} = \{0, 1\}$ (detection). This thesis is primarily interested in learning machines that describe either a class of probability densities or semi-parametric models which will be described shortly in this section.

Loss Functions

Once an appropiate model is chosen, it is typically fitted to the finite set of data $\mathbf{X}$ so that an estimate of the model's parameters θ can be calculated. To do so, practitioners leverage loss functions that establish a metric between the model f_θ and the data $\mathbf{X}$. A loss function is defined as

$$l : \Theta \times \mathcal{X} \mapsto \mathbb{R} : (\theta, \mathbf{x}) \longrightarrow l(\theta, \mathbf{x}), \tag{2.4}$$

where $\times$ denotes the Cartesian product. Later in this section, some loss functions that are specific to this thesis will be introduced.

Given a loss function l and an unknown data generating distribution $\mathcal{P}_0$, the underlying goal of statistical learning is to minimize the associated risk function $\mathcal{R} : \Theta \times \mathcal{P}_0 \mapsto \mathbb{R}$ to recover an optimal set of parameters θ^*, such that

$$\theta^* \coloneqq \arg\min_{\theta \in \Theta} \mathcal{R}\left(\theta, \mathcal{P}_0\right) \coloneqq \arg\min_{\theta \in \Theta} \mathbb{E}_{\mathbf{x} \sim \mathcal{P}_0} l\left(\theta, \mathbf{x}\right), \tag{2.5}$$

where $\mathbb{E}_{\mathbf{x} \sim \mathcal{P}_0}$ denotes the expectation over the true data distribution $\mathcal{P}_0$.

Empirical Risk Minimization

In practice, we do not know the true data generating distribution $\mathcal{P}_0$ and we only have access to the finite number of samples $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$. However, any dataset $\mathbf{X}$ can be represented by its associated empirical distribution

$$\mathcal{P}_N \coloneqq \frac{1}{N} \sum_{i=1}^{N} \delta_{\mathbf{x}_i}, \tag{2.6}$$

where $\delta_{\mathbf{x}}$ denotes the Dirac measure located at the point $\mathbf{x}$. Replacing $\mathcal{P}_\theta$ by its empirical counterpart $\mathcal{P}_N$ in Eqn 2.5, we get the associated empirical risk minimization

$$\hat{\theta} \;\coloneqq\; \arg\min_{\theta \in \Theta} \mathcal{R}\left(\theta, \mathcal{P}_N\right) \;=\; \arg\min_{\theta \in \Theta} \frac{1}{N} \sum_{i=1}^{N} l(\theta, \mathbf{x}_i). \tag{2.7}$$

This thesis concerns a particular class of learning machines. Chapters 5 and 6 utilise learning machines that define classes of probability densities, while Chapters 3 and 4 focus primarily on distribution-free semi-parametric models. We first define probability densities, however for further details on a host of different learning models see [8, 9].

2.1.1 Probability Density Fitting

A large class of learning models are parametrized directly by a probability density function (PDF). A PDF, denoted by $\mathcal{P}(x, \theta)$, is a function[1] whose value at any given sample $\mathbf{x}$ can be interpreted as the likelihood that the sample was drawn from the associated probability distribution. A PDF has the following key properties:

[1]The learning model described in Section 2.1 is equal to $f(\mathbf{x}, \theta) = \mathcal{P}(\mathbf{x}, \theta)$. However, we make the notation distinct to emphasis that the model is a density function.

- $0 \leq \mathcal{P}_\theta(\mathbf{x}) \leq 1$,

- $\int_{\mathcal{X}} \mathcal{P}_\theta(\mathbf{x}) d\mathbf{x} = 1$.

A popular example of a probabilistic model is that of a mixture model, which can be used to represent the presence of subpopulations within an overall population. Mixture models play an important role in Part II of this thesis where they are used to model the depth of surfaces in a lidar scene (see Section 2.5). A mixture model of K distributions is defined as

$$\mathcal{P}(\mathbf{x}; \theta) = \sum_{k=1}^{K} \alpha_k \mathcal{P}_k(\mathbf{x}; \theta_k), \tag{2.8}$$

where $\sum_{k=1}^{K} \alpha_k = 1$ and $\mathcal{P}_k(\mathbf{x}, \theta_k) \in \mathcal{P}(\mathcal{X})$ denotes the probability distribution of the kth mixture parameterized by θ_k. The parameters of the whole mixture model are denoted by $\theta = (\alpha_1, \ldots, \alpha_K, \theta_1, \ldots, \theta_K)$. In many mixture models, each of the k distributions $\mathcal{P}_k$ reduces to the same class of distributions such that $\mathcal{P}$ replaces $\mathcal{P}_k$ for each k in Eqn 2.8 for some consistent distribution $\mathcal{P}$. An instance of this is the Gaussian mixture model

$$\mathcal{P}(\mathbf{x}; \theta) = \sum_{k=1}^{K} \alpha_k \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{2.9}$$

where $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the multivariate Gaussian distribution over $\mathcal{X}$ parametrized by $\theta_k = (\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. As will be discussed in Section 2.3.3, Gaussian mixture models were one of the first learning models to be developed into a compressive learning scheme.

In general, given a dataset $\mathbf{X}$, one can infer the parameters θ of a probability distribution using several well-established techniques. The most classic is maximum likelihood estimation.

2.1.2 Maximum Likelihood Estimation

A traditional form of probability density fitting is maximum likelihood estimation (MLE). MLE attempts to estimate the true parameter θ_0 by maximizing the likelihood that the set of parameters generated the observations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$. Formally, the likelihood $L(\mathbf{X}; \theta)$ is defined as

$$L(\mathbf{X}; \theta) := \prod_{i=1}^{N} \mathcal{P}(\mathbf{x}_i; \theta). \tag{2.10}$$

In many scenarios, minimizing the log-likelihood $LL(\theta; \mathbf{X}) \coloneqq \log L(\mathbf{X}; \theta)$ defined as

$$LL(\mathbf{X}; \theta) \ \coloneqq \ \sum_{i=1}^{N} \log \mathcal{P}(\mathbf{x}_i; \theta) \tag{2.11}$$

is much easier in practice. The estimate $\hat{\theta}$ that minimizes the (log) likelihood function is called the maximum likelihood estimate. Notably, maximising Eqn 2.11 is equivalent to minimizing the empirical risk function in Eqn 2.7 where the loss function is set as $l(\theta, \mathbf{x}) = -\log \mathcal{P}_\theta(\mathbf{x})$.

For the Gaussian mixture model defined in Eqn 2.9, the log-likelihood function reduces to

$$LL(\mathbf{X}; \theta) \ = \ \sum_{i=1}^{N} \log \left(\sum_{k=1}^{K} \alpha_k \mathcal{N}\left(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \right), \tag{2.12}$$

which can be solved using the specific EM algorithm (see [10]).

One of the main disadvantages of maximum likelihood estimation (MLE) is that in some cases the likelihood function in Eqn 2.11 might not have a closed form solution nor a computationally tractable approximation. If that is the case, practitioners often resort to other probability density fitting techniques, including generalized method of moments.

2.1.3 Generalized Method of Moments

Generalized method of moments (GeMM) is an alternative form of parametric statistical inference where one estimates the parameters θ by matching a collection of generalized moments with an empirical counterpart computed over the observed samples $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$. Given a non-linear function $h : \mathbb{R}^d \mapsto \mathbb{C}^m$, then we define the expectation constraint of GeMM as

$$\mathbb{E}_{\mathbf{x} \sim \mathcal{P}_{\theta_0}} h(\mathbf{x}; \theta) = 0. \tag{2.13}$$

The non-linear function is chosen to ensure the parameters are identifiable i.e.

$$\mathbb{E}_{\mathbf{x} \sim \mathcal{P}_{\theta_0}} h(\mathbf{x}; \theta) = 0 \text{ if and only if } \theta = \theta_0. \tag{2.14}$$

The GeMM estimator is typically obtained by minimizing a quadratic loss of empirical discrep-

ancy with respect to θ to try impose the moment constraints of Eqn 2.13. Let us define

$$h_N(\mathbf{X};\theta) \;\coloneqq\; \frac{1}{N}\sum_{i=1}^{N} h(\mathbf{x}_i;\theta), \tag{2.15}$$

then a GeMM estimate classically takes the form

$$\hat{\theta} \coloneqq \arg\min_{\theta} h_N(\mathbf{X};\theta)^T \mathbf{W} h_N(\mathbf{X};\theta) \tag{2.16}$$

where $\mathbf{W} \in \mathbb{C}^{m\times m}$ is a symmetric positive definite (PD) weighting matrix that may depend on θ. It will discussed in Section 2.3 that generalised method of moments is similar to the concept of sketching, albeit with a fundamentally different goal.

2.1.3.1 Empirical Characteristic Function Estimation

Empirical characteristic function (ECF) is a specific class of GeMM that we build upon in Chapters 5 and 6 to construct a compact representation of the data. ECF has the particular separable form of h:

$$h(\mathbf{x};\theta) \;=\; \Phi(\mathbf{x}) \;-\; \mathbb{E}_{\mathbf{x}\sim\mathcal{P}_\theta}\Phi(x), \tag{2.17}$$

where $\Phi(\mathbf{x}) = \left[e^{\mathrm{i}\omega_j^T\mathbf{x}}\right]_{j=1:m}$ and ω_j are a discrete set of frequencies sampled from some law $\Lambda(\omega)$. They receive significant interest as the non-linear term is equal to the characteristic function in expectation i.e.

$$\Psi_{\mathcal{P}_\theta}(\omega) \;\coloneqq\; \mathbb{E}_{\mathbf{x}\sim\mathcal{P}_\theta}\Phi(x) \tag{2.18}$$

for the frequencies ω. The characteristic function has some unique and appealing properties, for instance, the characteristic function exists for all distributions and often has a closed form expression. The characteristic function captures all the information of the probability distribution, providing a one-to-one correspondence to the space of probability distributions. Moreover, under mild conditions, the characteristic function decays in frequency, i.e. $\Psi_{\mathcal{P}_\theta}(\omega) \to 0$ as $\omega \to \infty$.

2.1.4 Semi-Parametric Learning

So far we have considered parametric models $\mathcal{P}(\mathbf{x};\theta)$ that are parametrized fully by θ that is a vector in p-dimensional Euclidean space, i.e. $\Theta \subseteq \mathbb{R}^p$. In many learning tasks, however,

parametric models may be too restrictive as they require strong assumptions which may not be readily known. Often practitioners resort to non-parametric techniques, for instance k nearest neighbours, support vector machines and kernel regression (see [8]), that do not specify the model structure (i.e. a fixed number of parameters) a-priori and therefore the parameter set Θ is allowed to grow depending on the data. In contrast to parametric methods, the set of possible values of the parameter θ is a subset of some possibly infinite dimensional space V, such that $\Theta \subseteq V$ [11].

In this thesis, the class of semi-parametric models are considered. A semi-parametric model consists of both parametric and non-parametric terms. In this instance, the parameter set Θ is a subset of the cross product of both a finite and possibly infinite dimensional space [11, 12]:

$$\Theta \subseteq \mathbb{R}^p \times V. \tag{2.19}$$

Initially, it may seem that the class of semi-parametric models includes non-parametric models, however, it is often the case in semi-parametric learning that we are only interested in the finite dimensional component of θ. The possibly infinite dimensional component of θ that belongs to V is often referred to as the nuisance parameter [12].

In the following sections we consider two semi-parametric models, namely, principal component analysis (PCA) and independent component analysis (ICA) which are built upon in Chapters 3 and 4.

2.1.5 Principal Component Analysis

Principal component analysis (PCA) is a classic method in statistics, feature extraction and data compression where the goal is to project the data $\mathbf{X}$ on to an orthogonal coordinate system in which the basis vectors best describe the variability of the data.

Formally, PCA consists in finding the linear subspace of a fixed dimension $K < d$ that best fits the data $\mathbf{X}$ in the least squares sense. Assuming the data is centred, the goal of PCA is to find the K-dimensional orthogonal basis vectors (named principal vectors) $\mathbf{p}_1, \ldots, \mathbf{p}_K$ that maximises

$$\sum_{k=1}^{K} \sum_{i=1}^{N} |\mathbf{p}_k^T \mathbf{x}_i|^2. \tag{2.20}$$

There are many methods to finding the K principal vectors (see [8]). One of the most popular techniques consists of finding the K principal eigenvectors of the empirical covariance matrix $\Sigma_N = \frac{1}{N}\sum_{i=1} \mathbf{x}_i \mathbf{x}_i^T$ via the eigendecomposition:

$$\Sigma_N \simeq \mathbf{P}\mathbf{\Pi}\mathbf{P}^T \tag{2.21}$$

where the columns $\mathbf{p}_1, \ldots, \mathbf{p}_K$ of $\mathbf{P}$ are the K principal eigenvectors and the diagonal entries of $\mathbf{\Pi}$ are the corresponding eigenvalues.

In this instance, the principal vectors $\mathbf{p}_1, \ldots, \mathbf{p}_K$ correspond to the parametric components of Eqn 2.19 whilst the diagonal matrix $\mathbf{\Pi}$ represents the nuisance parameters defined in Eqn 2.19. It is therefore straightforward to see that PCA can be defined as a semi-parametric model.

2.1.6 Independent Component Analysis

Independent component analysis (ICA) is used frequently in the machine learning and signal processing communities to identify latent variables that are mutually independent to one another. It can be seen as an extension to PCA due to the assumption of independence between the latent variables which is stronger than the uncorrelated constraint in PCA.

To formulate the ICA problem, consider a data point $\mathbf{x} = (x_1, x_2, \ldots, x_d)^T$, then the problem of ICA concerns finding a mixing matrix $\mathbf{M} \in \mathbb{R}^{d \times n}$ such that

$$\mathbf{x} = \mathbf{M}\mathbf{s}, \tag{2.22}$$

where $\mathbf{s} = (s_1, s_2, \ldots, s_n)^T$ and the components s_i are statistically independent:

$$\mathcal{P}(s_1, s_2, \ldots, s_n) = \prod_{i=1}^{n} \mathcal{P}_i(s_i). \tag{2.23}$$

The following ambiguities in the ICA model in Eqn 2.22 hold:

- As both $\mathbf{s}$ and $\mathbf{M}$ are unknown, any scalar multiplier in one of the independent components s_j can always be cancelled by dividing the corresponding column in $\mathbf{M}$ by the same scalar.
- The order of the independent components and the corresponding columns in $\mathbf{M}$ can freely

change.

As a result, the mixing matrix $\mathbf{M}$ and the independent components $\mathbf{s}$ are only identifiable up to scaling and permutation ambiguities.

Prewhitening

A useful preprocessing strategy in ICA is to first whiten the observed variables $\mathbf{x}$ via a linear transformation

$$\mathbf{z} = \mathbf{V}\mathbf{x}, \tag{2.24}$$

such that the new vector $\mathbf{z}$ has identity covariance matrix. One popular method of whitening is to use the eigendecomposition of the covariance matrix (i.e. $\mathbb{E}_{\mathbf{x}\sim\mathcal{P}}\mathbf{x}\mathbf{x}^T$ as in Eqn 2.21 and set $\mathbf{V} = \mathbf{\Pi}^{-\frac{1}{2}}\mathbf{P}^T$. From Eqn 2.22 and 2.24 we have

$$\begin{aligned}\mathbf{z} &= \mathbf{\Pi}^{-\frac{1}{2}}\mathbf{P}^T\mathbf{M}\mathbf{s} \\ &= \mathbf{Q}\mathbf{s}.\end{aligned} \tag{2.25}$$

Whitening has two main advantages: (1) it handles the scenario when there are more mixing components than independent components $d > n$ as one can discard the $d - n$ smallest eigenvalues, (2) the matrix $\mathbf{Q} \in \mathbb{R}^{n\times n}$ in Eqn 2.25 is necessarily orthogonal and contains $n(n-1)/2$ degrees of freedom. The whitening matrix $\mathbf{V}$ is not unique and any orthogonal rotation of $\mathbf{V}$ will also define a whitening matrix [13]. For the sake of presentation, we will subsequently consider the whitened version of the data for the remainder of this section and the corresponding whitened ICA equation

$$\mathbf{z} = \mathbf{Q}\mathbf{s}. \tag{2.26}$$

There are many techniques and methods in the literature to solve the ICA problem. The simplest method is to assume the distributional form of each of the independent components $\mathcal{P}_i(s_i)$ and then solve the ICA problem through a maximum likelihood approach [14]. In practice, the distributions are not known a-priori so therefore in most methods the distributions are left unspecified. The unknown distributions $\mathcal{P}_i$ can be considered as the nuisance parameters of the semi-parametric model in Eqn 2.22, whilst the mixing matrix $\mathbf{M}$ corresponds to the finite parametric component. In Chapter 3, we develop a compressive learning framework for cumulant based ICA, therefore in this section we focus on such methods. For the interested reader, see

[13] for a exhaustive exposition on various ICA methods.

Cumulant Based ICA

Tensorial or cumulant based methods are a group of techniques used to solve the ICA problem that consist of using high order cumulant tensors. Tensors can be considered as a generalization of matrices or linear operators to higher order arrays. For instance, a 1st order tensor is a vector and a 2nd order tensor is a matrix. We denote a Kth order tensor by $\mathscr{T}^K \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_K}$, where the dimension of the tensors kth mode is denoted by I_k.

A cumulant tensor is a symmetric tensor whose components are functions of higher order moments of a random vector. Let $\mathscr{X}^K$ denote the Kth order cumulant tensor of a random variable $\mathbf{x} \in \mathbb{R}^d$, then its first four cumulant are defined [15] as

$$\begin{aligned}
\mathscr{X}^1_i &= \mathbb{E}[x_i] \\
\mathscr{X}^2_{ij} &= \mathbb{E}[x_i x_j] \\
\mathscr{X}^3_{ijk} &= \mathbb{E}[x_i x_j x_k] \\
\mathscr{X}^4_{ijkl} &= \mathbb{E}[x_i x_j x_k x_l] - \mathbb{E}[x_i x_j]\mathbb{E}[x_k x_l] - \mathbb{E}[x_i x_k]\mathbb{E}[x_j x_l] - \mathbb{E}[x_i x_l]\mathbb{E}[x_j x_k].
\end{aligned} \tag{2.27}$$

Given the whitened ICA model in Eqn 2.26 equating $\mathbf{z}$ to $\mathbf{s}$, then the following multilinear property holds for their associated cumulant tensors:

$$\mathscr{Z}^K = \mathscr{S}^K \times_1 \mathbf{Q} \times_2 \mathbf{Q} \times_3 \cdots \times_K \mathbf{Q}, \tag{2.28}$$

where $\times_j$ represents the j-mode tensor-matrix product and $\mathscr{S}^K$ represents the K^{th} order cumulant tensor of the independent components [15]. In Chapter 3, we will only consider 4^{th} order cumulant tensors (e.g. $K = 4$) and for the sake of simplified notation we shall drop the superscript in Eqn 2.28 for the rest of the discussion.

Formally, we can denote by $\mathfrak{C} \subset \mathbb{R}^{n \times n \times n \times n}$ the space of 4th order cumulant tensors which accounts for the symmetry in Eqn 2.27 and has a maximum of $\binom{n+3}{4}$ unique entries (degrees of freedom) [16]. The diagonal entries $\mathscr{Z}_{ijkl}$ $(ijkl = iiii)$ are the auto-cumulants of the whitened ICA data $\mathbf{z}$, while the off-diagonal entries $\mathscr{Z}_{ijkl}$ $(ijkl \neq iiii)$ are the cross-cumulants. An important property of cumulants is that if the variables are statistically independent then, as seen by Eqn 2.27, the cross-cumulants vanish to 0 resulting in a strictly diagonal cumulant

tensor. In other words, independence implies diagonality. It is shown in [17] that under mild conditions on the ICA model in Eqn 2.26 the converse is also true, i.e. diagonality implies independence. Once the data is whitened, the cumulant based ICA problem reduces to finding a linear transformation $\mathbf{Q}^T$ such that the resulting cumulant tensor

$$\mathscr{S} = \mathscr{Z} \times_1 \mathbf{Q}^T \times_2 \mathbf{Q}^T \times_3 \mathbf{Q}^T \times_4 \mathbf{Q}^T \tag{2.29}$$

is strictly diagonal.

The expected cumulant tensor $\mathscr{Z}$ is typically not known owing to finite data length approximations and non-Gaussian additive noise [18] and so in general $\mathscr{Z}$ cannot be *fully* diagonalized by a linear transform. As a result, contrast functions are used to approximately diagonalize $\mathscr{Z}$ and maximize the independence of the system.

Contrast Functions

Comon [19] proposed the use of contrast functions as a solution to tractably measure independence even when the independent components are left distribution-free. A contrast function $\varrho : \mathcal{P}(\mathcal{X}) \mapsto \mathbb{R}$ is a mapping from the space of distributions to the real line and can be thought of as a tractable approximation of mutual information. For a function ϱ to be a contrast function it must be both permutation and scale invariant, due to the ICA ambiguities, as well as being maximum if and only if components are statistically independent. Comon [19] proposed various cumulant based contrast functions that are Edgeworth expansions of information theoretic measures such as negative mutual information, maximum likelihood and negentropy. In Section 3.4, we utilise contrast functions as a measure of independence as part of our compressive ICA algorithms.

2.2 Estimation Theory

In Section 2.1, we described how parametric and semi-parametric models can be used to solve a learning task and different density estimation methods, for instance MLE and GeMM, were introduced. In this section, we discuss some of the key concepts of statistical inference and properties of an estimator $\hat{\theta}$. Many of these properties, for instance bias and efficiency, will be built upon in the latter chapters to analyse the proposed methods. As before, let

$\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N \overset{\text{i.i.d}}{\sim} \mathcal{P}_0$ be N i.i.d samples of an overarching probability distribution $\mathcal{P}_0$ that is parametrized by a set of true parameters $\theta_0 \in \Theta \subseteq \mathbb{R}^p$. Below we state some fundamental definitions of estimators.

2.2.1 Bias and Consistency

Definition 1 (Consistency)**.** *An estimate $\hat{\theta}$ is said to be consistent over the N samples $\mathbf{x}_1, \ldots, \mathbf{x}_N$ if and only if for all $\varepsilon > 0$*

$$\lim_{N \to \infty} \Pr\left(|\theta_0 - \hat{\theta}| > \varepsilon\right) = 0. \tag{2.30}$$

In other words, a consistent estimator is an estimator that's sampling distribution becomes increasingly concentrated around the true parameter θ_0 as the sample size N grows. Another important notion is the bias of an estimator.

Definition 2 (Bias)**.** *The bias of an estimator is defined as*

$$\textit{Bias}_{\theta_0}(\hat{\theta}) \coloneqq \mathbb{E}_{\mathbf{x} \sim \mathcal{P}} \hat{\theta} - \theta_0. \tag{2.31}$$

If $\textit{Bias}_{\theta_0}(\hat{\theta}) = 0$ then the estimator is said to be unbiased.

An unbiased estimator is one that in expectation equals the true parameter.

Initially, consistency and unbiasedness look similar, however an estimator can be consistent and biased while another estimator can be inconsistent yet unbiased. Take for instance an i.i.d sample $\mathbf{x}_1, \ldots, \mathbf{x}_N \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ and choose $\hat{\theta} = \mathbf{x}_N$ as the estimator for the true mean $\boldsymbol{\mu}$. Then by definition, $\mathbb{E}_{\mathbf{x} \sim \mathcal{P}} \mathbf{x}_N = \boldsymbol{\mu}$, however the estimator $\hat{\theta}$ doesn't converge towards a single value as $N \to \infty$ therefore $\hat{\theta}$ is unbiased yet inconsistent.

Under certain technical conditions, for instance a compact parameter space Θ and model identifiability, both the MLE and GeMM estimators discussed in Section 2.1.2 and 2.1.3, respectively, are consistent.

2.2.2 Central Limit Theorem

The central limit theorem (CLT) is a fundamental result in statistics and is built upon in Chapter 5 and 6 to develop the algorithms that will be proposed. It states that under certain conditions

the sum of i.i.d random variables converges to a Gaussian distribution. Let $\mathbf{x}_1, \ldots, \mathbf{x}_N$ be N i.i.d random variables and let $\bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i$, then formally the CLT states that

$$\sqrt{N}(\bar{\mathbf{x}} - \boldsymbol{\mu}) \xrightarrow{\text{dist}} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) \tag{2.32}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ denote the mean vector and covariance matrix of the random variables and $\xrightarrow{\text{dist}}$ denotes asymptotic convergence towards the CDF.

2.2.3 Efficiency

It may not be sufficient for an estimator to be consistent or unbiased as we may want $\hat{\theta}$ to converge to the true parameter θ_0 within a small amount of data. Such a performance can be measured by the efficiency of the estimator. Cramér and Rao proved a lower bound, named the Cramér-Rao lower bound (CRLB), that provides the optimal deviation in terms of variance that one can expect given an estimator $\hat{\theta}$. Specifically, the CRLB states [20] that for a unbiased[2] estimator

$$\text{Cov}\big(\hat{\theta}\big) \;=\; \Sigma_\theta \;\geq \mathcal{I}(\theta)^{-1}, \tag{2.33}$$

where $\mathcal{I}(\theta) \in \mathbb{R}^{p \times p}$ is the Fisher information matrix that has the entries

$$[\mathcal{I}(\theta)]_{ij} \;\coloneqq\; N\,\mathbb{E}_{\mathbf{x}\sim\mathcal{P}}\left[\left(\frac{\partial \log \mathcal{P}(\mathbf{x},\theta)}{\partial \theta_i}\frac{\partial \log \mathcal{P}(\mathbf{x},\theta)}{\partial \theta_j}\right)\right]. \tag{2.34}$$

In Eqn 2.33, the symbol $\geq$ refers to the matrix $\Sigma_\theta - \mathcal{I}(\theta)^{-1}$ being semi-positive definite. The efficiency of an estimator is a measure of how close an estimator is towards its CRLB. If an estimator $\hat{\theta}$ hits the CRLB (i.e. $\Sigma_\theta = \mathcal{I}(\theta)^{-1}$) then it is said to be fully efficient.

For the $k = 1$ case, the efficiency of an estimator $\hat{\theta}$ can be calculated by the ratio

$$e(\hat{\theta}) = \frac{1/\mathcal{I}_\theta}{\text{var}\left(\hat{\theta}\right)}, \tag{2.35}$$

where $0 \leq e(\hat{\theta}) \leq 1$ such that $e(\hat{\theta}) = 1$ results in a full efficient estimator. For an arbitrary parameter size p, the efficiency can be calculated using some discrepancy between Σ_θ and $\mathcal{I}_\theta$ which determines the loss of information incurred by using that estimator [21].

[2]There are similar results that give a CRLB for a biased estimator, see for example [21]

It is known that fully efficient estimators are rare and only exist if Θ defines an exponential family [21]. However, an estimator $\hat{\theta}$ is said to be asymptotically efficient if

$$\lim_{N\to\infty} \Sigma = \mathcal{I}_\theta^{-1}. \tag{2.36}$$

Figure 2.1 illustrates an example of an asymptotically efficient estimator.

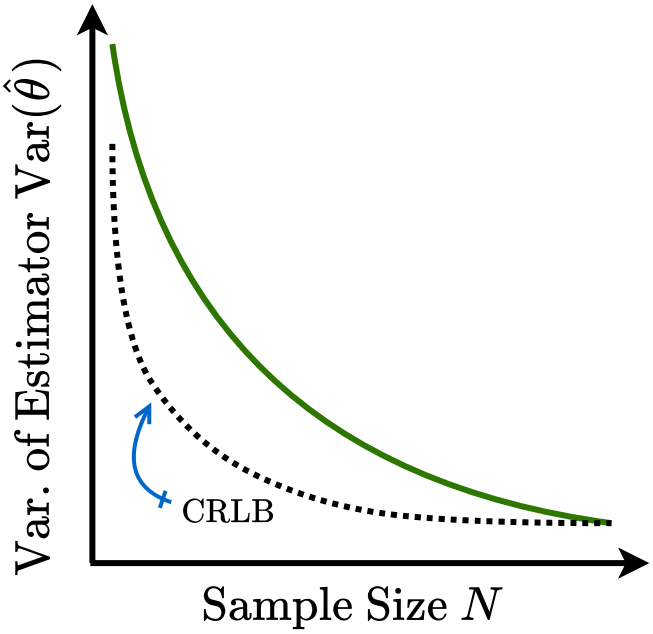

Figure 2.1: *An asymptotically efficient estimator achieves the CRLB as N becomes large.*

Under certain conditions, a MLE estimator, as discussed in Section 2.1.2, can be shown to be asymptotically efficient [21]. Furthermore, it can shown that a GeMM estimator is the most optimally efficient if the weighting matrix $\mathbf{W}$ is chosen such that [22, 23]

$$\mathbf{W} \propto \Omega_\theta^{-1} \tag{2.37}$$

where $\Omega_\theta \coloneqq \mathbb{E}_{\mathcal{X}\sim\mathcal{P}}\left[h_\theta(\mathbf{X};\theta)h_\theta(\mathbf{X};\theta)^T\right]$ is the covariance matrix of the non-linear function h_θ.

2.2.4 Sufficiency

Summary statistics are used to summarize a dataset $\mathbf{X} \in \mathbb{R}^{d\times N}$ in order to describe features of the data as simply as possible. The feature could be for example, the centre, deviation or shape of the data. For instance, one could take the empirical mean $\hat{\boldsymbol{\mu}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i$ or the empirical

covariance $\hat{\mathbf{\Sigma}} = \frac{1}{N-1}\sum_{i=1}^{N}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T$ to best summarize the data. Doing so would be an efficient manner of communicating the nature of the data and acts as a form of compression within its own right. If one was trying to fit a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$ to the dataset, then the empirical mean $\hat{\boldsymbol{\mu}}$ and covariance $\hat{\Sigma}$ would be sufficient in terms of retaining all the information needed to fit the model to the data. However, in general, simple summary statistics like the mean and covariance are not sufficient to fit an arbitrary distribution.

Sufficiency is an important notion in statistics and, as we will see in Section 2.3.2, is very related to the concept of forming a compact sketch. In its simplest form, a statistic is sufficient with respect to a family of distributions (parametric model) if no other statistic from the data would provide further information needed to estimate the parameters θ_0. The most trivial sufficient statistic of the data set $\mathbf{X}$ is itself $\mathbf{X}$. The Fisher–Neyman factorization theorem formalizes the idea of sufficiency [20]. Let $T(\mathbf{X})$ be a function over the individual samples $\mathbf{x}$ that computes a statistic (e.g. the empirical mean $T(\mathbf{X}) = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i$). Then $T(\mathbf{X})$ is called a sufficient statistic associated to the distribution $\mathcal{P}(\mathbf{X}, \theta)$ if and only if non-negative functions g_θ and $h_\mathcal{X}$ can be found such that the density function can be decomposed as

$$\mathcal{P}(\mathbf{X}, \theta) = h_\mathcal{X}(\mathbf{X}) g_\theta\left(T(\mathbf{X})\right). \tag{2.38}$$

In other words, the parameters of the distribution only interact with the data through the sufficient statistic $T(\mathbf{X})$. In fact, given any injective function λ then the statistic defined by the map $\lambda\left(T(\mathbf{X})\right)$ is also a sufficient statistic. Furthermore, a sufficient statistic $T(\mathbf{X})$ is a minimal sufficient statistic if, for any other sufficient statistic $T'(\mathbf{X})$, $T(\mathbf{X})$ is a function of $T'(\mathbf{X})$:

$$T(\mathbf{X}) \;=\; \lambda\left(T'(\mathbf{X})\right). \tag{2.39}$$

From a geometric perspective, this is equivalent to the maximum reduction without loss of information.

The Pitman-Koopman-Darmois theorem [24] states that sufficient statistics with bounded dimensionality exist only for distributions that belong to the exponential family. In other words, the size of the sufficient statistic $T(\mathbf{X})$ increases as the sample size N grows for non-exponential distributions. In the next section, we discuss the notion of compressive learning which attempts to construct approximate sufficient statistics that are of a fixed size relative to the size of the dataset.

2.3 Compressive Learning

In this section, the idea of compressive learning is introduced that is built upon in the following technical chapters. As discussed at the beginning of this chapter, the inspiration behind compressive learning originates from the field of compressive sensing where the foundations of theory and algorithm design are utilised in the principles of compressive learning in Section 2.3.2 as well as most of the following technical chapters. Details of the main concepts of compressive sensing are therefore first introduced below in Section 2.3.1 including the role of randomness, theoretical gaurantees and popular compressive sensing algorithms.

2.3.1 Compressive Sensing

Compressive sensing (CS) was developed in the early 2000s and introduced in the seminal papers of Candès et al. and Donoho [25, 26, 27]. The idea is to exploit the natural sparsity of signals to significantly reduce the sampling rate below the fundamental Shannon-Nyquist threshold rate and save on the complexities associate with sensing.

Let $\mathbf{A} \in \mathbb{R}^{m \times d}$ be linear measurement operator, then the standard CS problem can be defined as

$$\mathbf{y} = \mathbf{A}\mathbf{x}_0 + \mathbf{e}, \tag{2.40}$$

where the goal is to recover the original signal $\mathbf{x}_0 \in \mathbb{R}^d$ from a set of measurements $\mathbf{y} \in \mathbb{R}^m$ that have been corrupted by some noise $\mathbf{e} \in \mathbb{R}^m$. In CS the measurement operator $\mathbf{A}$ is often dimension reducing (i.e. $m < d$) therefore the number of measurements is typically smaller than the signal dimension. In such a case, the recovery of $\mathbf{x}_0$ from limited measurements is theoretically ill-posed, even in the noiseless case. To make the problem defined in Eqn 2.40 well-posed and tractable, one must introduce regularity assumptions, for instance sparsity, on the signal set.

Role of Sparsity

Recovery of the true signal in Eqn 2.40 is possible if the true signal is sparse, meaning only a few elements of the signal are non-zero. Many signals of interest are indeed sparse within some appropiate domain or basis, for instance, natural images are sparse in the wavelet domain due

to the strong correlation between neighbouring pixels. Formally, a signal $\mathbf{x} \in \mathbb{R}^d$ is k-sparse if

$$\|\mathbf{x}\|_0 \;\leq\; k, \tag{2.41}$$

where $\|\cdot\|_0 \coloneqq |\text{Supp}(\mathbf{x})|$ denotes the ℓ_0-norm which counts the number of non-zero elements of $\mathbf{x}$. For simplicity, we assume that the signal $\mathbf{x}$ is sparse in its canonical basis of $\mathbb{R}^d$.

As it will be shown later in this section, a signal $\mathbf{x}_0$ can be recovered if we assume it (approximately) belongs to a low-dimensional model set $\mathfrak{S}_\theta$. In the case of standard CS, the low-dimensional model set can be defined as

$$\mathfrak{S}_\theta \;=\; \mathfrak{S}_k \;\coloneqq\; \{\mathbf{x} \mid \|\mathbf{x}\|_0 \leq k\}\,. \tag{2.42}$$

Role of Randomness

Given the true signal $\mathbf{x}_0 \in \mathfrak{S}_k$ is k-sparse, a key question is how can one design measurement operators $\mathbf{A}$ that can recover the true signal $\mathbf{x}_0$? Candès and Tao introduced an important tool in CS called the restricted isometry property (RIP) that, if satisfied, ensures the measurement operator $\mathbf{A}$ provides a stable embedding [28] for signals in the model set.

Definition 3 (Restricted Isometry property)**.** *A measurement matrix $\mathbf{A} \in \mathbb{R}^{m\times d}$ satisfies the RIP on a set $\mathfrak{S}_k$ with constant δ if $\forall\, \mathbf{x}, \mathbf{x}' \in \mathfrak{S}_k$*

$$(1-\delta)\|\mathbf{x}-\mathbf{x}'\|_2^2 \leq \|\mathbf{A}(\mathbf{x}-\mathbf{x}')\|_2^2 \leq (1+\delta)\|\mathbf{x}-\mathbf{x}'\|_2^2. \tag{2.43}$$

From a geometric perspective, as demonstrated in Figure 2.2, the RIP ensures that the measurement operator preserves pointwise distances between members of the model set.

Do linear measurement operators that satisfy the RIP exist? Interestingly, they exist for a class of operators that are randomly distributed, i.e. $\mathbf{A}_{ij} \sim \Lambda$, where the distributing law Λ satisfies the following concentration inequalities:

$$\mathbb{E}_{\mathbf{A}_{ij}\sim\Lambda}\left(\|\mathbf{A}\mathbf{x}\|^2\right) \;=\; \|\mathbf{x}\|^2, \tag{2.44}$$

and

$$\Pr\left(\left|\|\mathbf{A}\mathbf{x}\|^2 - \|\mathbf{x}\|^2\right| \geq \epsilon\|\mathbf{x}\|^2\right) \;\leq\; 2e^{-dc_0} \tag{2.45}$$

for a constant c_0 depending on ϵ. Under suitable additional conditions on m and k, many simple distributions satisfy these concentration inequalities, including subgaussian and symmetric Bernoulli distributions. In standard CS, it is shown in [29] (see Theorem 1.4) that measurement operators that are randomly generated from the Gaussian distribution $\mathbf{A}_{ij} \sim \mathcal{N}(0, m^{-\frac{1}{2}})$ satisfy the RIP in Eqn 2.43 with high probability provided that

$$m \geq Ck \log\left(\frac{d}{k}\right) \tag{2.46}$$

for some constant C. Next, we detail some classical recovery algorithms in the CS literature which we build upon in later chapters.

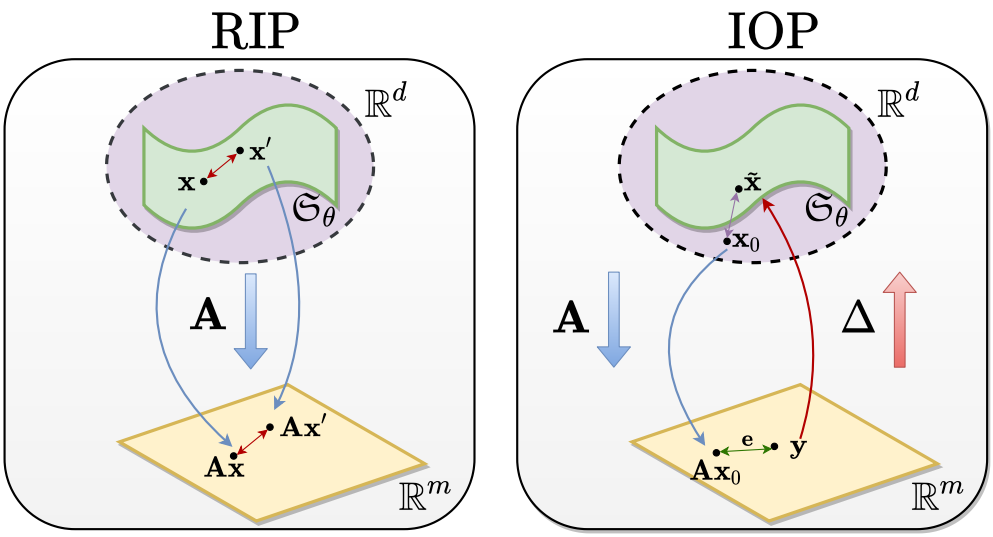

Figure 2.2: *Schematic of the restricted isometry property (left) and the instance optimal decoder property (right).*

Recovery Algorithms

Theoretically, one could recover the k-sparse signal $\mathbf{x}_0$ by solving the following ℓ_0 optimization problem

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2 \qquad \text{s.t} \quad \|\mathbf{x}\|_0 \leq k. \tag{2.47}$$

Intuitively, Eqn 2.47 would return the k-sparse vector $\mathbf{x}$ that minimises the measurement error $\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2$. However, it is well known that the ℓ_0 optimization problem in Eqn 2.47 is N-P hard due to the combinatorics involved in checking the support of $\mathbf{x} \in \mathbb{R}^d$ for large d [30].

Many algorithms have been designed to solve the CS problem in Eqn 2.47. Here we provide details of convex relaxation methods and greedy approaches. The greedy approach of iterative projected gradient is of particular interest as it is the main inspiration behind the compressive ICA algorithm in Chapter 3.

Convex relaxation of Eqn 2.47 can be achieved by replacing the ℓ_0 constraint with an ℓ_1 constraint, resulting in the following ℓ_1 optimization problem:

$$\hat{\mathbf{x}} \;=\; \arg\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2 \qquad \text{s.t} \quad \|\mathbf{x}\|_1 \leq k'. \tag{2.48}$$

The ℓ_1-norm induces sparse solutions as shown in Figure 2.3. Under the form of Eqn 2.48, the problem is known as the LASSO but it is equivalent to other optimization problems such as basis pursuit denoising or constrained ℓ_1 minimization [31]. As Eqn 2.48 is convex, is it both tractable and efficient to solve.

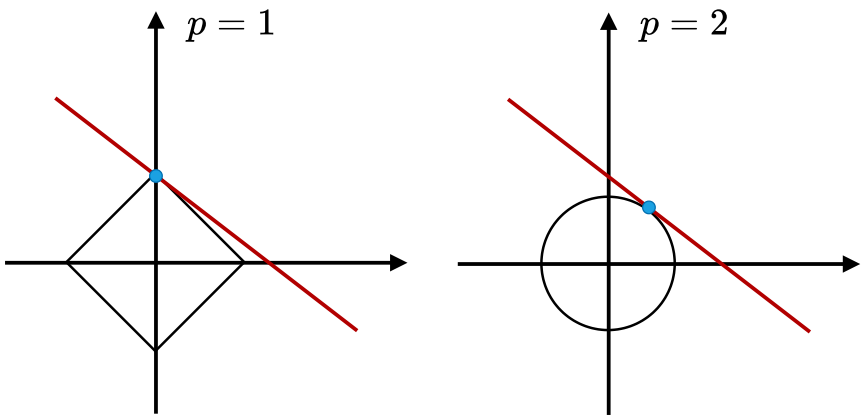

Figure 2.3: *ℓ_p norms for $p = 1, 2$ in 2D. The red line represents the linear constraints induced by the observations. A solution of the ℓ_p problem occurs when the linear constraints intersects with the ℓ_p ball. The ℓ_1 problem induces a sparse solution i.e. the intersection occurs on the axis.*

Greedy algorithms are also used to solve the standard CS problem in Eqn 2.47. The term *greedy* is used in the sense that these methods make the locally optimal choice at each iteration of the algorithm. Iterative projected gradient (IPG) is a popular constrained optimization method which enforces the regularity assumptions imposed by the model set $\mathfrak{S}_\theta$ by projecting the object on to the model set $\mathfrak{S}_\theta$ after each subsequent gradient step [32]. An IPG scheme can be defined

by the following recursive steps:

$$\textbf{Iterative Projected Gradient}-\text{Initialize } \mathbf{x}^0 \in \mathbb{R}^d$$
$$\text{For} \quad j = 0, 1, 2, ..., K$$
$$\left\lfloor \begin{array}{l} \text{Gradient Step: } \mathbf{x}^{j+\frac{1}{2}} = \mathbf{x}^j + \mu_j \nabla_{\mathbf{x}} l \\ \text{Projection Step: } \mathbf{x}^{j+1} = \mathcal{P}_{\mathfrak{S}_\theta}\left(\mathbf{x}^{j+\frac{1}{2}}\right) \end{array} \right.$$

where $\nabla_{\mathbf{x}} l$ denotes the gradient of the loss function respect to $\mathbf{x}$ and where $\mathcal{P}_{\mathfrak{S}_\theta}$ defines the (orthogonal) projection operator defined as

$$\mathcal{P}_{\mathfrak{S}_\theta}(\mathbf{x}^*) \in \underset{\mathbf{x} \in \mathfrak{S}_\theta}{\arg\min} \|\mathbf{x}^* - \mathbf{x}\|. \tag{2.49}$$

By orthogonal we refer to the projection operator $\mathcal{P}_{\mathfrak{S}_\theta}$ projecting the point $\mathbf{x}^*$ on to the point on the model set that is closest with respect to some norm $\|\cdot\|$. For the case of k-sparse vectors, Blumensath et al. [33] proposed the iterative hard thresholding (IHT) algorithm that projects $\mathbf{x}^*$ onto the k-sparse model set $\mathfrak{S}_k$ by thresholding the $d - k$ smallest absolute entries of $\mathbf{x}^*$ to zero.

Orthogonal matching pursuit is another instance of a greedy algorithm that is used frequently in CS. At each iteration, the OMP algorithm selects the columns of $\mathbf{A}$ which are most correlated to the current residual error $\mathbf{r}^j = \mathbf{x}^j - \mathbf{x}$. The column is then added into the support set. At each iteration, the algorithm updates the residual $\mathbf{r}^j$ by projecting the measurements $\mathbf{y}$ onto the linear subspace spanned by the columns currently selected in the support set. Compared to other greedy algorithms, OMP is simple and fast as it only requires k iterations to approximate a k-sparse signal. See [34] for a thorough exposition on greedy algorithms used in CS.

Generalization to Low-Dimensional Spaces

Compressive sensing began with roots situated in sparsity. However, CS can be applied to other signals that are assumed to reside in some low-dimensional model set $\mathfrak{S}_\theta$. For example, low-rank matrix recovery and the related work of matrix completion can be formulated as a CS problem:

$$\hat{\mathbf{X}} = \underset{\mathbf{X} \in \mathfrak{S}_r}{\arg\min} \|\mathbf{y} - \mathcal{A}(\mathbf{X})\|, \tag{2.50}$$

where the measurement operator $\mathcal{A} : \mathbb{R}^{d\times d} \mapsto \mathbb{R}^m$ is defined as $\mathcal{A}(\cdot) \coloneqq \mathbf{A}\text{vec}(\cdot)$, where vec denotes the vectorization operator. In this instance, the low-dimensional model set[3] is defined as

$$\mathfrak{S}_\theta \;=\; \mathfrak{S}_r \;\coloneqq\; \{\mathbf{X} \mid \text{rank}(\mathbf{X}) \leq r\}. \tag{2.51}$$

Similar to the LASSO problem in Eqn 2.48, the low-rank matrix recovery problem can be solved via the following optimization problem:

$$\hat{\mathbf{X}} \;=\; \underset{\mathbf{X}\in\mathfrak{S}_r}{\arg\min} \|\mathbf{y} - \mathcal{A}(\mathbf{X})\|_2^2 \qquad \text{s.t} \quad \|\mathbf{X}\|_* \leq r'. \tag{2.52}$$

where $\|\cdot\|_*$ denotes the nuclear norm which acts a tractable surrogate used to minimize the rank of a matrix. Random matrices that satisfy the concentration inequalities in Eqn 2.44 and Eqn 2.45 satisfy the RIP over the model set $\mathfrak{S}_r$ of rank-r matrices provided that [35]

$$m \;\geq\; Crd. \tag{2.53}$$

As a result, substantial compression can be attained if the rank $r \ll d$.

CS has also been applied to signals that belong to other generic low-dimensional spaces including manifolds [36]. In most cases, the size of the projected space m needs only to scale with the dimensionality of the normalized secant set of $\mathfrak{S}_\theta$, defined as

$$\mathfrak{N}(\mathfrak{S}_\theta - \mathfrak{S}_\theta) \;\coloneqq\; \left\{ \frac{\mathbf{x}_1 - \mathbf{x}_2}{\|\mathbf{x}_1 - \mathbf{x}_2\|} \mid \mathbf{x}_1 \neq \mathbf{x}_2 \in \mathfrak{S}_\theta \right\}, \tag{2.54}$$

such that

$$m \gtrsim \dim\left(\mathfrak{N}(\mathfrak{S}_\theta - \mathfrak{S}_\theta)\right) \tag{2.55}$$

for some measure of dimensionality dim. One classical measure of dimensionality is that of the upper-box counting dimension (UBCD) denoted $\dim_{\text{B}}$ that is defined as

$$\dim_{\text{B}} \;\coloneqq\; \lim_{\epsilon\to 0} \sup \log\left[\text{CN}(\mathfrak{S}_\theta, \|\cdot\|, \epsilon)\right] / \log\left[1/\epsilon\right]. \tag{2.56}$$

Here $\text{CN}(\mathfrak{S}_\theta, \|\cdot\|, \epsilon)$ defines the covering number of the normalized secant set that defines the minimum number of closed balls of radius ϵ, with respect to the norm $\|\cdot\|$, with centres in $\mathfrak{N}(\mathfrak{S}_\theta - \mathfrak{S}_\theta)$ that are needed to cover $\mathfrak{N}(\mathfrak{S}_\theta - \mathfrak{S}_\theta)$ [31, 37]. The set of centres that cover

[3] For simplicity we assume the matrices are square, however this still holds for rectangular matrices $\mathbf{X} \in \mathbb{R}^{d_1\times d_2}$.

$\mathfrak{N}\left(\mathfrak{S}_\theta - \mathfrak{S}_\theta\right)$ is called the ϵ-net [31]. In many well-behaved cases, the UBCD of the normalized secant set is twice that of the model set $\mathfrak{S}_\theta$ [37]. The notion of covering numbers and ϵ nets will be built open in Chapter 3 when developing our main theoretical result. We next provide theoretical gaurantees on recovering a signal that belongs to generic low-dimensional model set $\mathfrak{S}_\theta$.

Robust Decoding

A decoder Δ is an algorithm, such as the ones discussed previously, that approximately solves the CS problem associated with a low-dimensional model set $\mathfrak{S}_\theta$. Formally, the decoder $\Delta : \left(\mathbb{R}^m, \mathbb{R}^m \times \mathbb{R}^d\right) \mapsto \mathbb{R}^d$ recovers an estimate $\hat{\mathbf{x}}$ such that

$$\hat{\mathbf{x}} \;=\; \Delta\left(\mathbf{y}, \mathbf{A}\right). \tag{2.57}$$

The ultimate goal of CS is to recover an estimate $\hat{\mathbf{x}}$ from (noisy) measurements $\mathbf{y}$ that are approximately close to the true signal $\mathbf{x}_0$. Bourrier et al. [38] established a connection between the lower RIP (i.e. the LHS of Eqn 2.43) and the existence of a so-called instance optimal decoder Δ such that the decoder estimate $\hat{\mathbf{x}} = \Delta\left(\mathbf{y}, \mathbf{A}\right)$ satisfies the following instance optimal property (IOP):

$$\|\hat{\mathbf{x}} - \mathbf{x}_0\| \;\leq\; \alpha d\left(\mathbf{x}_0, \mathfrak{S}_\theta\right) \;+\; \beta\|\mathbf{e}\| \tag{2.58}$$

for some norm $\|\cdot\|$, constants $\alpha, \beta \geq 0$ and where $d\left(\mathbf{x}, \mathfrak{S}_\theta\right)$ denotes the distance between $\mathbf{x}$ and the model set $\mathfrak{S}_\theta$ defined as

$$d\left(\mathbf{x}, \mathfrak{S}_\theta\right) \;\coloneqq\; \inf_{\mathbf{x}_{\mathfrak{S}_\theta} \in \mathfrak{S}_\theta} d(\mathbf{x}, \mathbf{x}_{\mathfrak{S}_\theta}), \tag{2.59}$$

for some distance metric d. Fundamentally, the IOP states that the error of the instance optimal decoder is bounded by the modelling error $d\left(\mathbf{x}_0, \mathfrak{S}_\theta\right)$ and the measurement error $\mathbf{e}$. Furthermore, the IOP shows that the error caused by both model mismatch and sampling error has controlled amplification as of result of using the measurement operator $\mathbf{A}$.

It is shown in [38] that if the measurement operator $\mathbf{A}$ satisfies the lower RIP for the model set $\mathfrak{S}_\theta$ with constant $(1 - \delta)$, then the decoder defined by

$$\Delta\left(\mathbf{y}, \mathbf{A}\right) \;=\; \arg\min_{\mathbf{x} \in \mathfrak{S}_\theta} \|\mathbf{y} - \mathbf{A}\mathbf{x}\| \tag{2.60}$$

satisfies the IOP in Eqn 2.58 for the measurement operator $\mathbf{A}$ and the model set $\mathfrak{S}_\theta$ with constants $\alpha = 1$ and $\beta = 2(1-\delta)^{-1}$. Subsequently, the pair $(\mathbf{A}, \Delta)$ provide a tractable CS scheme given that $\mathbf{A}$ satisfies the RIP on the low-dimensional model set $\mathfrak{S}_\theta$ as demonstrated by Figure 2.2. This result forms the basis of the main information preservation theorem of compressive ICA introduced in Chapter 3.

2.3.2 Principles of Compressive Learning

By setting out the disciplines that inspired compressive learning in Sections 2.1.1 and 2.3.1, it is now possible to introduce the main principles of compressive learning which are then built upon in the following chapters. The core principles of compressive learning can be summarized by the schematic in Figure 2.4 where the compressive learning process is split into 2 phases, namely, the sketching and learning phase. We begin this section by setting out the goal of compressive learning and discussing in detail the initial sketching phase.

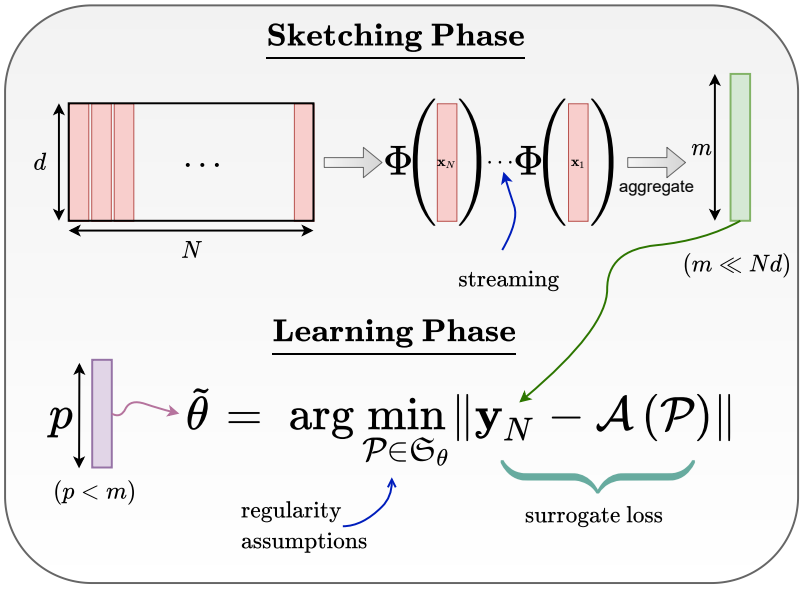

Figure 2.4: *The core principles of compressive learning split into the sketching phase and the learning phase.*

2.3.2.1 Sketching Phase

The original compressive learning (CL) framework was developed in [6, 39] by Gribonval, Keriven and coauthors with the goal of estimating the parameters θ_0 of some true probability distribution $\mathcal{P}_0 \in \mathcal{P}(\mathcal{X})$ from only limited measurements. The main inspiration was to general-

ize the CS framework, discussed in Section 2.3.1, by compressing any distribution $\mathcal{P} \in \mathcal{P}(\mathcal{X})$ via a linear measurement operator $\mathcal{A} : \mathcal{P}(\mathcal{X}) \mapsto \mathbb{C}^m$. As will become clear, the compressed representation $\mathcal{A}(\mathcal{P})$, or so-called sketch, of the probability distribution $\mathcal{P}$ can enable substantial reductions in the complexities associated with parameter inference such as memory and computation as discussed in Section 2.1.

The linear measurement operator $\mathcal{A} : \mathcal{P}(\mathcal{X}) \mapsto \mathbb{C}^m$, which in the context of compressive learning we call a sketching operator, is defined as

$$\mathcal{A}(\mathcal{P}) \coloneqq \mathbb{E}_{\mathbf{x}\sim\mathcal{P}}\Phi(\mathbf{x}) \tag{2.61}$$

where $\Phi : \mathcal{X} \mapsto \mathbb{C}^m$ is a non-linear feature function associated with the sketching operator $\mathcal{A}$. The resulting sketch $\mathbf{y} = \mathcal{A}(\mathcal{P})$ can interpreted as a collection of generalized moments of the probability distribution $\mathcal{P}$. It is important to notice that, although the feature function Φ associated with the sketch is a non-linear, the sketching operator $\mathcal{A}$ is in fact linear over the space of probability distributions, for instance

$$\mathcal{A}(\alpha\mathcal{P}_1 + (1-\alpha)\mathcal{P}_2) = \alpha\mathcal{A}(\mathcal{P}_1) + (1-\alpha)\mathcal{A}(\mathcal{P}_2). \tag{2.62}$$

for $\mathcal{P}_1, \mathcal{P}_2 \in \mathcal{P}(\mathcal{X})$ and $\alpha \in \mathbb{R}$.

As discussed in Section 2.1.1, one does not have access to the true probability distribution but to a finite set of i.i.d observations $\mathbf{x}_1, \ldots, \mathbf{x}_N \overset{\text{i.i.d}}{\sim} \mathcal{P}_0$. The empirical sketch, denoted by $\mathbf{y}_N$ is then computed as

$$\mathbf{y}_N \coloneqq \mathcal{A}(\mathcal{P}_N) = \frac{1}{N}\sum_{i=1}^{N}\Phi(\mathbf{x}_i). \tag{2.63}$$

In comparison to the CS problem in Eqn 2.40, the goal of CL is to recover the true distribution $\mathcal{P}_0$ from

$$\mathbf{y}_N = \mathcal{A}(\mathcal{P}_0) + \mathbf{e} \tag{2.64}$$

where the error $\mathbf{e} \coloneqq \mathcal{A}(\mathcal{P}_0) - \mathcal{A}(\mathcal{P}_N)$ converges to zero as N tends to infinity by the law of large numbers.

The empirical sketch $\mathbf{y}_N$ can be computed extremely efficiently over the dataset $\mathbf{X}$ as only a

single pass of each observation is required and the dataset does not need to be stored in memory thereafter. Furthermore, as $\mathbf{y}_N$ is simply an empirical average, the computation of the sketch is easily parallelizable and is very suited to both data streams and distributed learning (see [40]). In Chapter 5, it will be shown that these properties make a sketch extremely appealing when constructing a compact representation of a lidar dataset on chip. As will be discussed next, all that is required to estimate the parameters of the probability distribution is the compact sketch. Fundamentally, the size of the sketch m does not scale with the dimensions of the dataset $\mathbf{X}$, but is instead driven by the complexity or dimensionality of the parameter space Θ associated with the probability distribution of interest. We again draw the analogy back to CS in Section 2.3.1 whereby the number of measurements required scales with the underlying sparsity of the true signal $\mathbf{x}_0$. From a parameter inference perspective, the sketch can be interpreted as constructing an approximate sufficient statistic to the probability distribution as the size of the sketch m is set fixed and is independent of the length N of the dataset. We next turn our attention to the learning stage of CL as illustrated in Figure 2.4.

2.3.2.2 Learning Phase

Once the sketch is computed as in Eqn 2.63, one must learn the parameters of the distribution solely from the empirical sketch $\mathbf{y}_N$. To that end, we minimize a task specific cost function $C(\theta, \mathbf{y}_N)$ to infer an estimate $\tilde{\theta}$ such that

$$\tilde{\theta} \coloneqq \underset{\theta \in \Theta}{\arg\min}\, C(\theta, \mathbf{y}_N). \tag{2.65}$$

The cost function $C(\theta, \mathbf{y}_N)$ acts as a surrogate to the risk function $\mathcal{R}(\theta, \mathcal{P}_0)$, discussed in Section 2.1, to the extent that the sketch estimate $\tilde{\theta}$ is close to the ERM estimate $\hat{\theta}$ in some sense. In many compressive learning schemes (see Section 2.3.3), the cost function reduces to some distance between the empirical sketch $\mathbf{y}_N$ and the associated (expected) sketch, for instance

$$\tilde{\theta} = \underset{\mathcal{P} \in \mathfrak{S}_\theta}{\arg\min} \|\mathbf{y}_N - \mathcal{A}(\mathcal{P})\|, \tag{2.66}$$

for some norm $\|\cdot\|$. This is reminiscent of the signal recovery problem of CS in Eqn 2.40 where we seek to recover a sparse signal from limited measurements. In a compressive sensing light one can introduce regularity assumptions to make Eqn 2.66 well-posed. These regularity

assumptions come in the form of a low-dimensional model set that solutions to the learning task lie on or a close to. Formally, a model set $\mathfrak{S}_\theta$ of a learning task is defined as

$$\mathfrak{S}_\theta \coloneqq \{\mathcal{P} \in \mathcal{P}(\mathcal{X}) \mid \exists \theta \in \Theta, \mathcal{R}(\theta, \mathcal{P}) = 0\}. \tag{2.67}$$

In other words, the model set defines the set containing all distributions over $\mathcal{P}(\mathcal{X})$ for which zero risk is achievable.

Similar to CS, one can design a decoder Δ that exploits the structural assumptions of the model set that takes as inputs the empirical sketch $\mathbf{y}_N$ and the sketching operator $\mathcal{A}$ and recovers the parameters θ of the model, e.g.

$$\tilde{\theta} \;=\; \Delta(\mathbf{y}_N, \mathcal{A}).$$

2.3.2.3 Sketch Design

Up until now, the specific design of a sketch has not be discussed. Gribonval et al. [1, 6] proposed the use of random Fourier features (RFF) where the feature function is defined as:

$$\Phi(\mathbf{x}) \;=\; \frac{1}{\sqrt{m}}\left[e^{\mathrm{i}\omega_j^t \mathbf{x}}\right]_{j=1:m} \tag{2.68}$$

where $\omega_1, \ldots, \omega_m$ are m frequencies sampled i.i.d from some distributing law Λ. RFFs have the favourable property that in expectation they are equal to the characteristic function:

$$\mathcal{A}(\mathcal{P}) \;=\; \mathbb{E}_{\mathbf{x}\sim\mathcal{P}}\Phi(\mathbf{x}) \;=\; \frac{1}{\sqrt{m}}\left[\Psi_{\mathcal{P}}(\omega_j)\right]_{j=1:m}. \tag{2.69}$$

As discussed in Section 2.1.3.1, characteristic functions exist for all distributions and in many cases they have a closed form well-behaved expression, making the optimization problem in Eqn 2.66 particularly tractable.

Remark 1. *When using RFF based sketching operators, CL is similar to ECF estimation which was introduced in Section 2.1.3.1. However, CL is distinctively different in its broader goal: ECF aims to estimate parameters θ of a distribution when the likelihood is unavailable, while CL attempts to estimate θ by building, in theory, the most compact representation of the data as possible. Nonetheless, we can leverage well-established theory and algorithms from the ECF literature to better understand RFF based CL.*

It turns out that the RFF sketching operator in Eqn 2.69 is very related to finite kernel approximations and kernel mean embedding of distributions. Before we draw the connection, let us briefly introduce the definitions of reproducing kernel Hilbert spaces and kernels.

A kernel $\kappa : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{C}$ defines a similarity measure between points $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$. The Moore-Aronszojn [41] theorem states that a positive definite[4] (PD) kernel κ is associated to a unique Hilbert space $\mathcal{H}_\kappa$ that satisfies the following properties: for any $\mathbf{x} \in \mathcal{X}$ the function $\kappa(\mathbf{x}, \cdot)$ belong to $\mathcal{H}_\kappa$, and the kernel satisfies the reproducing property $\forall f \in \mathcal{H}_\kappa, \forall \mathbf{x} \in \mathcal{X}$ then $f(\mathbf{x}) = \langle f(\cdot), \kappa(\mathbf{x}, \cdot) \rangle_{\mathcal{H}_\kappa}$. The space $\mathcal{H}_\kappa$ is referred to as the reproducing kernel Hilbert space (RKHS) and $\langle \ \rangle_{\mathcal{H}_\kappa}$ denotes the inner product defined in $\mathcal{H}_\kappa$. We refer the reader to [42] for a comprehensive review on kernels and RKHS.

In [1, 6], the authors proposed constructing the RFF sketch in Eqn 2.68 by sampling the frequencies from a Gaussian distribution $\Lambda = \mathcal{N}(\mathbf{0}, \sigma^{-2}\mathbf{I}_d)$. The inner product $\langle \Phi(\mathbf{x}), \Phi(\mathbf{x}') \rangle \simeq \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}'\|_2^2}{2\sigma^2}\right)$ approximates a Gaussian kernel. In general, to design an appropiate sampling distribution Λ, one can leverage Bochner's theorem [43] that allows one to write any PD translation invariant kernel (i.e. $\kappa(\mathbf{x}, \mathbf{x}') = \kappa(\mathbf{x} - \mathbf{x}')$) as the inverse Fourier transform of a probability distribution i.e.

$$\kappa(\mathbf{x} - \mathbf{x}') \ = \ \int_{\mathbb{R}^d} e^{\mathrm{i}\omega^T(\mathbf{x}-\mathbf{x}')} \, \mathrm{d}\Lambda(\omega) \ = \ \mathbb{E}_{\omega \sim \Lambda} e^{\mathrm{i}\omega^T(\mathbf{x}-\mathbf{x}')}. \tag{2.70}$$

A direct consequence of Bochner's theorem is that the inner product of the RFF denoted $\kappa_\Phi(\mathbf{x}', \mathbf{x}) = \langle \Phi(\mathbf{x}), \Phi(\mathbf{x}) \rangle$ approaches the true kernel in expectation. Given an appropiate kernel, one can design a sampling strategy by sampling the frequencies from the Fourier transform of the kernel Λ. In general, the choice of kernel is often a difficult task in itself and may require cross validation techniques. In Chapter 5, we use both a random sampling scheme as described above as well as deterministic sampling scheme to construct a sketch of the lidar data.

The RFF sketching operator defined in Eqn 2.69 is also deeply related to kernel mean embeddings of distributions. The kernel mean embedding denoted by $\mu_\mathcal{P} : \mathcal{P}(\mathcal{X}) \mapsto \mathcal{H}_\kappa$, defined as

$$\mu_\mathcal{P}(\cdot) \ \coloneqq \ \int_\mathcal{X} \kappa(\mathbf{x}, \cdot) \, \mathrm{d}\mathcal{P}(\mathcal{X}) \ = \ \mathbb{E}_{\mathbf{x} \sim \mathcal{P}} \, \kappa(\mathbf{x}, \cdot) \tag{2.71}$$

provides a map from the space of distributions to the RKHS associated with the PD kernel κ. Of

[4] A positive definite kernel is one which the associated kernel Gram matrix $\mathbf{G}_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j)$ is positive definite (see [42]).

particular interest is the class of characteristic kernels which are PD and have the extra property that the kernel mean embedding $\mu_{\mathcal{P}}$ is injective, that is to say, there is a one-to-oneness between the space of distributions and the RKHS. The injectiveness allows us to define a metric over the RKHS that measures the distance between distributions $\mathcal{P}, \mathcal{Q} \in \mathcal{P}(\mathcal{X})$:

$$\text{MMD}(\mathcal{H}_\kappa, \mathcal{P}, \mathcal{Q})^2 \coloneqq \|\mu_{\mathcal{P}}(\cdot) - \mu_{\mathcal{Q}}(\cdot)\|^2_{\mathcal{H}_\kappa}. \tag{2.72}$$

The distance, denoted by MMD, is called the maximum mean discrepancy and can be expressed explicitly [44] in terms of the associated kernel κ:

$$\text{MMD}(\mathcal{H}_\kappa, \mathcal{P}, \mathcal{Q})^2 = \mathbb{E}_{\substack{\mathbf{x}\sim\mathcal{P} \\ \mathbf{x}'\sim\mathcal{P}}} \kappa(\mathbf{x}, \mathbf{x}') - 2\,\mathbb{E}_{\substack{\mathbf{x}\sim\mathcal{P} \\ \mathbf{y}\sim\mathcal{Q}}} \kappa(\mathbf{x}, \mathbf{y}) + \mathbb{E}_{\substack{\mathbf{y}\sim\mathcal{Q} \\ \mathbf{y}'\sim\mathcal{Q}}} \kappa(\mathbf{y}, \mathbf{y}'). \tag{2.73}$$

Interestingly, using Bochner's theorem from Eqn 2.70, we can also express the MMD in the Fourier domain using the characteristic functions of the distribution $\mathcal{P}, \mathcal{Q}$, denoted $\Psi_{\mathcal{P}}$ and $\Psi_{\mathcal{Q}}$, respectively:

$$\text{MMD}(\mathcal{H}_\kappa, \mathcal{P}, \mathcal{Q})^2 = \int |\Psi_{\mathcal{P}}(\omega) - \Psi_{\mathcal{Q}}(\omega)|^2 \, \mathrm{d}\Lambda(\omega). \tag{2.74}$$

As a result, the RFF sketch in Eqn 2.68 can be interpreted as a finite-dimensional approximation of the kernel mean embedding in Eqn 2.71.

Aside from the choice of feature function, the number of frequencies drawn, and therefore the size of the sketch m, plays a critical role in the construction of the sketch. Ideally, one would like m be large enough to capture enough salient information required to estimate the parameters θ of the model. A general principle stemming from the CS framework in Section 2.3.1 states that the size of the sketch needs to be approximately $m = \mathcal{O}(|\mathfrak{S}_\theta|)$. For the case of CL, the dimensionality of the model set equates to the size of the parameter space Θ i.e. the number of parameters p. To that end, we typically require $m \gtrsim p$. However, it will be shown in this thesis that the efficiency (see Section 2.2.3), and therefore the loss of information incurred by the sketch estimator $\tilde{\theta}$ is determined by the size of m.

2.3.2.4 Learning Guarantees of Compressive Learning

In a similar nature to CS, we seek a sketching operator $\mathcal{A}$ that is stable to both sampling noise and modelling error. The lower RIP, as introduced in Section 2.3.1, is also a principal tool in CL. In order to measure the distance between probability distributions with respect to a learning task, we introduce the risk induced metric $\|\cdot\|_{\mathcal{R}}$ that is defined as

$$\|\mathcal{P} - \mathcal{Q}\|_{\mathcal{R}} \coloneqq \sup_{\theta \in \Theta} |\mathcal{R}(\theta, \mathcal{P}) - \mathcal{R}(\theta, \mathcal{Q})|. \tag{2.75}$$

The lower RIP in CL therefore states that for some constant $\delta \geq 0$ and for $\mathcal{P}, \mathcal{Q} \in \mathfrak{S}_\theta$

$$(1-\delta)\|\mathcal{P} - \mathcal{Q}\|_{\mathcal{R}}^2 \leq \|\mathcal{A}(\mathcal{P}) - \mathcal{A}(\mathcal{Q})\|_2^2. \tag{2.76}$$

In other words, the lower RIP states that an appropiate sketching operator preserves the relative distance between distributions with respect to a risk induced metric. Gribonval et al. [1, 6] proved that if a sketching operator $\mathcal{A}$ satisfies the lower RIP in Eqn 2.76 then the excess risk is controlled. For instance, with high probability

$$\mathcal{R}\left(\tilde{\theta}, \mathcal{P}_0\right) - \mathcal{R}(\theta^*, \mathcal{P}_0) \leq d(\mathcal{P}_0, \mathfrak{S}_\theta) + \beta \|\mathbf{y}_N - \mathcal{A}(\mathcal{P}_0)\|_2 \tag{2.77}$$

for some constant β and where θ^* denotes the minimizer of the true risk and $d(\mathcal{P}_0, \mathfrak{S}_\theta)$ denotes some distance from $\mathcal{P}_0$ to the nearest point on the model set. Note that if the sketching operator satisfies the lower RIP then the excess risk is bounded by modelling error and sampling noise.

Remark 2. *The initial CL guarantees assume that the model $\mathcal{P}_0$ is parametric. However, in many cases, for instance semi-parametric models, the model is left only partially specified and is typically identified through some non-parametric statistic of the data (see Section 2.1.4). In Chapter 4, we elaborate on the difficulties of building a CL scheme for semi-parametric models.*

2.3.3 Existing Compressive Learning Models

Keriven et al. [45] pioneered the fist CL models that centred around the task of mixture modelling. Recall from Eqn 2.8, that a mixture model $\mathcal{P}(\mathbf{x}; \theta)$ is defined as

$$\mathcal{P}(\mathbf{x}; \theta) = \sum_{k=1}^{K} \alpha_k \mathcal{P}_k(\mathbf{x}; \theta_k) \tag{2.78}$$

where $\mathcal{P}_k$ denotes the probability distribution of the kth mixture parameterized by θ_k. Equation 2.78 promotes a low-dimensional model set by restricting a priori a maximum of K mixtures that can model the data. Keriven considered both the K-means mixture model, where $\mathcal{P}_k(\mathbf{x};\theta_k) = \delta_{\mathbf{c}_k}$ and $\theta_k = \mathbf{c}_k$, as well as the Gaussian mixture model where $\mathcal{P}_k(\mathbf{x};\theta_k) = \mathcal{N}(\mu_k, \mathbf{\Sigma}_k)$ and $\theta_k = (\mu_k, \mathbf{\Sigma}_k)$[5]. The authors proposed using RFFs as introduced in Eqn 2.68 to form the sketch. Advantageously, the expected sketch is equal to the linear combination of each mixtures characteristic function, i.e.

$$\mathbb{E}_{\mathbf{x}\sim\mathcal{P}}\Phi(\mathbf{x}) \;=\; \Psi_{\mathcal{P}}(\omega) \;=\; \sum_{k=1}^{K}\alpha_k\Psi_{\mathcal{P}_k}(\omega) \tag{2.79}$$

due to the linearity of characteristic functions. The CL cost function reduces to

$$\tilde{\theta} \;:=\; \underset{\theta\in\Theta}{\arg\min}\|\mathbf{y}_N - \sum_{k=1}^{K}\alpha_k\Psi_{\mathcal{P}_k}(\omega)\|_2^2. \tag{2.80}$$

Conveniently, the characteristic functions are known in advance where $\Psi_{\delta_\mathbf{c}}(\omega) = e^{\mathrm{i}\omega^T\mathbf{c}}$ and $\Psi_{\mathcal{N}}(\omega) = e^{\mathrm{i}\omega^T\mu-\frac{1}{2}\omega^T\mathbf{\Sigma}\omega}$ are the characteristic functions for the Dirac and Gaussian distribution, respectively. In [46], the authors proposed the CLOMP algorithm which is a greedy, matching pursuit inspired algorithm (see Section 2.3.1) which iteratively extends the support of the parameter set by choosing at each step the atom which is most correlated with the residual error. It can be shown [1] that a sketch of size $m \gtrsim K^2d$ satisfies the CL lower RIP in Eqn 2.76.

Aside from compressive mixture models, Gribonval et al. introduced a compressive PCA scheme in [6] which has strong connections to low-rank matrix recovery. Recall from Section 2.1.5 that the PCA model attempts to find a K-dimensional linear subspace that describes the maximal variance of the data. The PCA problem can be solved by finding the K largest eigenvalues associated to the covariance matrix $\Sigma_\theta = \mathbb{E}_{\mathbf{x}\sim\mathcal{P}}\mathbf{x}\mathbf{x}^T$. The model set $\mathfrak{S}_\theta$ is defined as

$$\mathfrak{S}_\theta \;=\; \{\mathbf{\Sigma}_\theta \mid \mathrm{rank}\,(\mathbf{\Sigma}_\theta) \leq K\} \tag{2.81}$$

which enforces the low-dimension regularity assumptions. In this instance, Gribonval [6] pro-

[5] Keriven considered the case when the covariance matrices were isotropic e.g. $\mathbf{\Sigma}_k = \sigma_k^2\mathbf{I}_d$

posed the sketching operator defined by

$$\mathbf{y}_N \;=\; \mathcal{A}\left(\boldsymbol{\Sigma}_\theta\right) \;=\; \mathbf{A}\text{vec}\left(\boldsymbol{\Sigma}_\theta\right) \;=\; \frac{1}{N}\sum_{i=1}^{N}\langle \mathbf{a}_j, \mathbf{x}_i\mathbf{x}_i^T\rangle_{j=1:m} \tag{2.82}$$

where $\mathbf{a}_j$ is the jth column of a random matrix $\mathbf{A}$. The CL cost function reduces down

$$\tilde{\boldsymbol{\Sigma}} \;\coloneqq\; \underset{\boldsymbol{\Sigma}\in\mathfrak{S}_\theta}{\arg\min}\|\mathbf{y}_N - \mathcal{A}\left(\boldsymbol{\Sigma}\right)\|_2^2 \tag{2.83}$$

which can solved using the nuclear norm formulation of the low-rank matrix recovery in Section 2.3.1. It can be shown that a sketch of size $m \gtrsim Kd$ is sufficient to satisfy the lower RIP.

Remark 3. *As introduced in Section 2.1.5, the PCA model is semi-parametric due to the fact the distributional form $\mathbf{x} \sim \mathcal{P}$ is typically left unspecified. By its very nature, we leverage the covariance statistic $\boldsymbol{\Sigma}_\theta$ to identify the parameters θ of the PCA model. As a result, the CL PCA model reduces to a finite dimensional compressive sensing problem. In Chapter 4, we develop the concept of semi-parametric compressive learning further.*

2.3.4 Advantages and Disadvantages of CL

Compressive learning partially addresses some of the challenges of large-scale learning. Below we state some of the unique advantages and disadvantages of CL.

- The salient information required to infer the parameters θ of a model can be typically captured by a sketch of size $m = \mathcal{O}(k)$. Crucially, the size of the sketch only scales linearly with respect to the complexity of the learning task and, fundamentally, is independent of the dimensions of a dataset $\mathbf{X}$.
- The sketching operator $\mathcal{A}$ is linear with respect to the probability distribution one is attempting to recover. As a result, the sketch is tremendously parallelizable, for instance, one can aggregate local sketches from decentralized devices or servers to form a *global* sketch. This can significantly reduce the data transfer requirements of sending batches of high dimensional data. Moreover, the linearity of the sketching operator allows one to compute the sketch on the fly in real time which is important in the context of streaming data as we will see in Part II of the thesis.
- The parameters θ of the model are inferred solely from the sketch and therefore poten-

tially sensitive data does not need to be transferred or stored locally where it is most vulnerable to malicious attacks. Moreover, due to the random construction of the sketch, it produces a natural encryption of the data. The privacy of the sketch can be further increased by adding i.i.d noise to the sketch in its construction, where the level of variance of noise determining the overall privacy level. See the works of Chatalic et al. [47] for the detailed works on the privacy gaurantees of compressive learning.

- The parameters of the model are optimised by minimizing a task-specific compressive learning cost function that acts as a proxy to the risk. As long as the sketching operator is chosen appropriately, such that it satisfies the RIP condition, then CL is robust with respect to modelling error and sampling noise at recovering a estimate $\tilde{\theta}$ that exhibits a controlled level of excess risk.

- Depending on the distribution or model $\mathcal{P}$ that one is trying to recover, the expected sketch $\mathcal{A}(\mathcal{P}) = \mathbb{E}_{\mathbf{x}\sim\mathcal{P}}\Phi(\mathbf{x})$ may not have a closed form expression. This can be quite problematic when attempting to minimise the cost function in Eqn 2.65 as defining the gradient with respect to the parameters θ would be difficult.

- As part of a compressive learning algorithm, we are required to promote the regularity assumptions of the model set $\mathfrak{S}_\theta$ to accurately recover an estimate of the true parameters. This often requires either a tractable regularizer or a projection operator, as discussed in Section 2.3.1, which may be difficult or intractable to implement. This will be discussed in more detail in Chapter 3, where we develop a compressive ICA algorithm.

2.4 Large Scale Learning

In recent years, both the length of the datasets N and the size d have substantially increased due part to the advances in technology (i.e. more sensors) and the emergence of automated processes that can record increasingly amounts of data. It is commonplace to have datasets $\mathbf{X}$ where $N \sim$ millions and $d \sim$ thousands requiring up to terrabytes of storage memory. The size of such large datasets can cause severe challenges as $\mathcal{O}(Nd)$ space complexity is required on local RAM while the computational complexity can equate to days/weeks of runtime. To combat the complexities associated with large scale learning, many methods have been proposed to approximate the dataset in some particular manner. Although not exhaustive, the current approaches can be grouped into 3 main categories: sub-sampling, projections and sketches,

depicted in Figure 2.5.

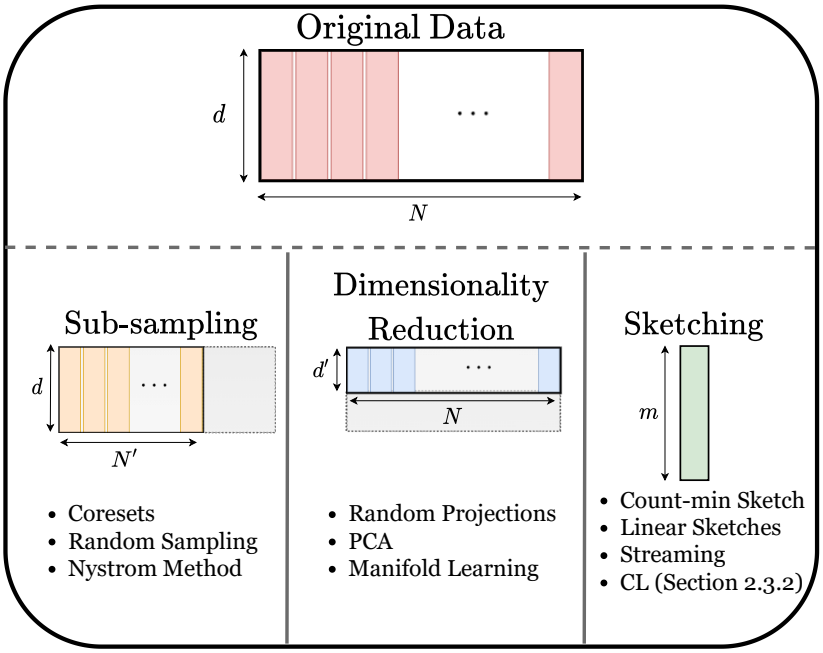

Figure 2.5: *The three main approaches used to tackle the issues surrounding large scale learning.*

In this section, the aforementioned groups of methods will be introduced while discussing their advantages and limitations.

2.4.1 Sub-Sampling

Sub-sampling is a class of methods that attempts to reduce the complexities of memory and/or computation associated with the length of the data N. The principal idea is to form a subset $\mathbf{X}' \in \mathbb{R}^{d \times N'}$ of the original data $\mathbf{X}$ such that

$$\mathbf{X}' \coloneqq \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{N'}\}, \quad \mathbf{x} \in \mathbb{R}^d, \tag{2.84}$$

where $N' \leq N$. Sub-sampling methods can be categorized into adaptive and non-adaptive schemes with respect to the data.

Non-adaptive Sub-sampling

Non-adaptive sub-sampling techniques are typically simple and cheap methods that are used to form a subset $\mathbf{X}'$ of the data. By non-adaptive, we refer to the host of sub-sampling techniques that don't change their behaviour relative to different datasets. Random sampling is the simplest non-adaptive sub-sampling method that draws a sample of size N' from $\mathbf{X}$ by repeating the following steps N' times:

Random Sub-Sampling$-$For $j = 0, 1, 2, ..., N'$

1) Produce a random number from 1 to N: $r_j \sim \Lambda(N)$

2) Obtain the r_j^{th} item in $\mathbf{X}$ (i.e. $\mathbf{x}_{r_j}$) and add it to the sub-sample $\mathbf{X}'$

Many random distributions $\Lambda(N)$ have been proposed, however the most established is uniform sampling such that $r \sim \mathcal{U}(1, N)$ where $\mathcal{U}(a, b)$ is the discrete uniform distribution defined over the interval (a, b). Once a sample is drawn, one can either replace the sample back into the population, in this case $\mathbf{X}$, or sample without replacement. However, for large N and small N' the difference between the two diminishes as the probability of drawing the same integer more than once is relatively low. As random sampling is a naive non-adaptive form of sub-sampling, there is a risk that important information is discarded during the sampling process. This is a particular concern in classification tasks where there are unrepresented groups (i.e. unbalanced classes). Poisson sampling and stratified sampling can be used in this scenario to give a larger weight to certain classes or partitions of the data to increase the probability of sampling.

Coresets

In recent years, more advanced sub-sampling techniques have been developed which are data dependent: of these is coresets. Coresets optimally find a subset $\mathbf{X}'$ of a dataset $\mathbf{X}$ by optimising over a certain measure associated to the set [48]. Formally, a subset $\mathbf{X}'$ is called an ϵ-coreset of $\mathbf{X}$ with respect to a measure function μ if

$$(1 - \epsilon)\mu\left(\mathbf{X}'\right) \;\leq\; \mu\left(\mathbf{X}\right). \tag{2.85}$$

Coresets are usually task specific and are therefore heavily reliant to the measure μ that is used. They have had success in the tasks of clustering including k-means [40]. One of the major limitations with coresets is that to form a ϵ-coreset, the size of the coreset must scale with

$$N' \;=\; \mathcal{O}\left(1/\epsilon\right). \tag{2.86}$$

Moreover, in practice coresets are limited to so-called *faithful* measures, e.g. diameter, width and smallest enclosing ball, as the computation complexity of computing the coreset only has a linear dependency in N. However, non-faithful measures typically exhibit a larger polynomial dependency in N making their use impractical.

Nyström Method

The Nyström method is a popular sub-sampling technique used primarily in kernel methods such as support vector machines, kernel regression and kernel PCA (see [42]). Typically in kernel methods, a Gram matrix $\mathbf{G} \in \mathbb{C}^{N\times N}$, defined as $\mathbf{G}_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j)$, associated with the data matrix $\mathbf{X} \in \mathbb{R}^{d\times N}$ has to be stored and computed resulting in a memory and computational cost of $\mathcal{O}\left(N^2\right)$ and $\mathcal{O}\left(N^3\right)$, respectively. For large N this quickly becomes infeasible. Let I denote the set of indices that correspond to N' sampled columns of $\mathbf{G}$. Then the Nyström method [49, 50] attempts to find an approximate Gram matrix $\tilde{\mathbf{G}}$ such that

$$\tilde{\mathbf{G}} \;\coloneqq\; \mathbf{G}_I \mathbf{G}_{I,I}^{-1} \mathbf{G}_I^T \tag{2.87}$$

where $\mathbf{G}_I \in \mathbb{R}^{N\times N'}$ is the matrix containing the sampled columns of $\mathbf{G}$ and $\mathbf{G}_{I,I} \in \mathbb{R}^{N'\times N'}$ is a submatrix of $\mathbf{G}$ obtained by further subsampling the rows of $\mathbf{G}_I$. As a result, the space and computational complexity is reduced to $\mathcal{O}\left(NN'\right)$ and $\mathcal{O}\left(NN'^2\right)$, respectively, as one does not need to store the original kernel Gram matrix. The Nyström method can be seen as an adaptive low-rank approximation of the Gram matrix, and therefore the number of samples N' required is related to the rank r of $\mathbf{G}$.

In practice, the rank of the Gram matrix is far less than the length of the dataset (i.e. $r \ll N$), therefore the compression achieved by the Nyström method can be substantial, as reported on real datasets in [49].

A major drawback with all sub-sampling techniques is that they do not tackle the complexities

associated with size of the data d. Machine learning tasks can therefore remain infeasible if d is significantly large.

2.4.2 Dimensionality Reduction

Dimensionality reductions are used frequently to reduce the complexities associated with the size of the data d. The idea revolves around mapping the data to a smaller dimensional space via a map $\Phi : \mathbb{R}^d \mapsto \mathbb{R}^{d'}$ where $d' \leq d$. Dimensionality reduction can be grouped into random or structured projections.

Random Projections

The Johnson-Lindenstrauss (J-L) lemma stated below is an important result that is closely related to compressive sensing (see Section 2.3.1).

Lemma 1 (Johnson-Lindenstrauss Lemma [51])**.** *Let $d' \geq \mathcal{O}\left(\epsilon^{-2} \log N\right)$ be an integer. For every set $\mathbf{X}$ of N points on $\mathbb{R}^d$, there exists a linear mapping $\Phi : \mathbb{R}^d \mapsto \mathbb{R}^{d'}$ such that for all $\mathbf{x}_1, \mathbf{x}_2 \in \mathbf{X}$*

$$(1-\epsilon)\|\mathbf{x}_1 - \mathbf{x}_2\|^2 \;\leq\; \|\Phi(\mathbf{x}_1) - \Phi(\mathbf{x}_2)\|^2 \;\leq\; (1+\epsilon)\|\mathbf{x}_1 - \mathbf{x}_2\|^2. \tag{2.88}$$

Importantly, without any prior knowledge of the data, the J-L lemma states that there exists a linear function $\Phi : \mathbb{R}^d \mapsto \mathbb{R}^{d'}$ that satisfies the RIP condition (see Section 2.3.1) provided $d' \gtrsim \epsilon^{-2} \log N$. Although the J-L lemma is an existence result, it has been shown [52] that maps consisting of random matrices i.e.

$$\Phi(\mathbf{x}) \;=\; \mathbf{A}\mathbf{x} \quad \text{for } \mathbf{A} \in \mathbb{R}^{d' \times d} \sim \Lambda \tag{2.89}$$

satisfy the J-L lemma. Of those is the subgaussian matrix that was introduced in Eqn 2.3.1. However, in practice, fast Johnson Lindenstrauss transforms (FJLT) are used as a computationally efficient alternative [53]. FJLTs are constructed by the product of simple matrices that are computationally quick to compute, for example Hadamard and sparse matrices. The computational complexity can be shown to reduce from $\mathcal{O}(dd')$ to $\mathcal{O}(d \log d')$ by replacing the dense subgaussian matrices with a FJLT.

Structured Projections

Structured projections are an alternative dimensionality reduction tool that, in contrast to random projections, are data-dependent. PCA, as introduced in Section 2.1.5, is a linear structured projection that is widely used as a preprocessing tool to reduce the dimensions of the data. PCA projects the data onto a linear subspace that represents the maximal variance of the data. In many cases, a high proportion of the data's variance is captured within the first few principal components (we often observe an exponential decay in practice). Substantial compression can therefore be attained by projecting the data subspace spanned by the first d' principal vectors. Other linear subspace learning models, for example ICA, linear discriminant analysis (LDA) and canonical correlation analysis can also be recast as a dimensionality reduction technique, however, they are used less ofen in practice.

Non-linear structured projections are an alternative approach, used particularly for complex data, to reduce size of the feature space d. Kernel PCA can be used by generalising standard PCA to non-linear dimension reduction by projecting the data to non-linear subspaces of maximal variance. Locally linear embedding (LLE) is another approach that seeks a low dimension projection that preserves local neighbourhood distance of the data with respect to some measure. The idea is to capture the possible manifold structure of the data. Other manifold learning based dimensionality reduction include isomaps, t-SNE and autoencoder networks. See [54] for further details on non-linear projections that are used in dimensionality reduction. In general, non-linear based approaches to dimensionality reduction can exhibit a high computational complexity and in some instances may be more challenging to solve than the original machine learning task.

Both random and structured projections only partially alleviate the complexities of fitting a model to some large dataset $\mathbf{X}$ as the length N of the dataset remains unaltered. Dimensionality reduction techniques are therefore only feasible in certain scenarios when the length of the data is fairly moderate.

2.4.3 Sketching

In the world of drawing, the method of sketching serves as a simple, quickly made illustration that records what the artist perceives as important in the scene. Sketching from a learning perspective is no different. Although the term *sketching* can vary in definition depending on

the field, it shares the core value of computing a sketch that is a fixed size representation of a dataset $\mathbf{X}$ that is built specifically for a particular model or task. In this thesis, a sketch $\mathbf{y} \in \mathbb{C}^m$ is defined as

$$\mathbf{y} \;=\; \mathcal{F}\left(\mathbf{X}\right) \in \mathbb{C}^m \tag{2.90}$$

for some operator $\mathcal{F}$.

Historically, sketches stemmed from the field of approximate query processing (AQP) where short descriptors of a large dataset were constructed such that they could approximately answer certain queries [40]. For instance, the count-min sketch [55] serves the purpose of detecting items in a database that frequently appear, named *heavy hitters*. However, the biggest driver behind recent sketch development is that they are particularly easy to update. This is important in AQP as relational databases are constantly being updated with a stream of data and therefore a sketch must be efficient to update so that the answer to a query can remain accurate. For that reason, the class of linear sketches are of particular interest. Linear sketches have the additional property that

$$\mathbf{y} \;=\; \mathcal{F}\left(\mathbf{X}\right) \;=\; \sum_{i=1}^{N} \tilde{\Phi}(\mathbf{x}_i) \tag{2.91}$$

for some linear function $\tilde{\Phi} : \mathbb{R}^d \mapsto \mathbb{C}^m$. As the sketch update requires only an operation followed by a summation, they are extremely amenable to the streaming context where computational runtime is paramount. Furthermore, a *global* sketch can be easily constructed through the aggregation of local sketches making linear sketches easily adaptable to distributed learning where a model is fitted over multiple decentralized devices or servers.

Sketches are not just reserved for relational databases. Tropp et al. showed in [56] that linear sketches can be used to approximate low-rank matrices, that are too large to be stored in memory, in the streaming context. In many large scale scientific simulations [57], including climate forecasting, fluid dynamics and aircraft design, a low-rank matrix $\mathbf{H} \in \mathbb{R}^{N\times d}$ can be decomposed into a sequence

$$\mathbf{H} \;=\; \mathbf{H}_1 \;+\; \mathbf{H}_2 \;+\; \mathbf{H}_3 \;+\; \ldots$$

where each innovation $\mathbf{H}_i$ is assumed to be low-rank (or have other redundancies like sparsity or structure). In many scenarios, the innovation $\mathbf{H}_i$ is processed and then discarded due to memory constraints. In [56], the authors used FJLTs as part of the linear sketch such that

$\tilde{\Phi}(\mathbf{x}) = \mathbf{A}\mathbf{x}$. Due to the linearity of the sketch, one can form a sketch of $\mathbf{H}$ by taking recursive sketches of each innovation $\mathbf{H}_i$ e.g.

$$\mathbf{y} \;=\; \mathcal{F}\left(\mathbf{H}\right) \;=\; \tilde{\Phi}(\mathbf{H}_1) \;+\; \tilde{\Phi}(\mathbf{H}_2) \;+\; \tilde{\Phi}(\mathbf{H}_2) \;+\ldots$$

The resulting sketch $\mathbf{y}$ can be then used to recover a low-rank matrix to approximate $\mathbf{H}$, as discussed in Section 2.3.1. One of the disadvantages of this method is that one might not know a priori the approximate rank of the matrix $\mathbf{H}$ so selecting an appropiate sketch size m can be difficult. Moreover, several passes of the data may be required to reduce the low-rank approximation error.

Compressive learning, as introduced in Section 2.3.2, falls under the category of sketching. However, there is a subtle yet important difference compared to the (linear) sketching techniques discussed in this section. The compressive learning sketch defined in Eqn 2.61 is linear over the space of distributions $\mathcal{P}(\mathcal{X})$ but typically non-linear over the data. In contrast, the sketches defined in this section are linear over the dataset.

2.5 Single Photon Counting Lidar

In Part II of this thesis, we will apply the ideas of CL introduced in Section 2.3.2 to the depth imaging modality of single photon light detection and ranging (lidar) that currently suffers from generating an excess of data. As will be discussed shortly in this section, the excess of data causes a data transfer bottleneck on lidar devices which can lead to suboptimal depth images. Below we state the fundamentals and challenges of single photon lidar that are built upon in PArt II.

Single photon lidar has emerged as an important depth imaging technique prevalent in the defense [58], forestry [59] and automobile industries [60, 61]. In contrast to other depth imaging modalities, lidar has the advantage of offering very high depth resolution even at long range scenes using low power, eye safe lasers. At the core of the technique is the ability of emitting light pulses and detecting each single photon as it arrives, thereby obtaining a depth estimate by measuring the round-trip time of individual photons for each pixel in the scene.

Figure 2.7 depicts a schematic of a typical time-correlated single photon counting (TCSPC) lidar system. A laser emits a pulse wave of photons to a scene that triggers the system clock

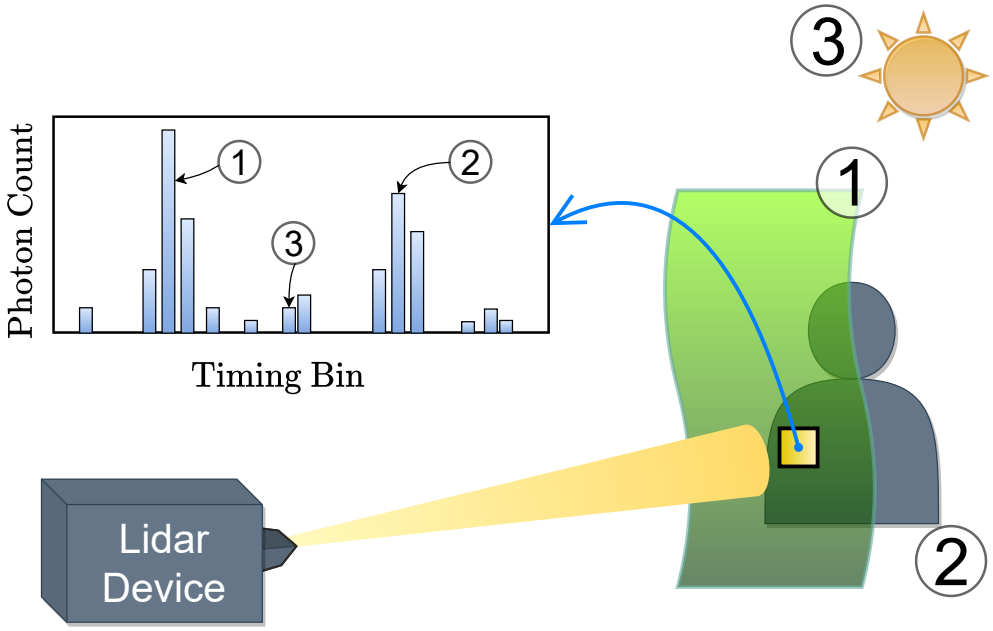

Figure 2.6: *Single photon lidar: a laser is directed at a pixel in a scene which consists of a semi-transparent camouflage netting with a human stood behind. The recorded photons produce a TCSPC histogram that exhibits 2 peaks of varying intensity for the camouflage surface (1) and human (2), as well as spurious photon detections from ambient sources like the sun (3).*

where a single photon avalanche diode (SPAD) is used to detect individual photons. Specifically, the SPAD consists of a reverse-biased photodiode which, in the presence of a photon, induces an avalanche of electrical charge carriers that are directly detectable as a digital signal. A time-to-digital converter (TDC) then converts the signal to a digital time-stamp that updates a timing statistic in an online manner. Here the *timing statistic* terminology refers to the various traditional methods that are used to collect and store the time-stamps. The most commonly used timing statistic is that of a TCSPC histogram, as depicted in Figure 2.6, that clusters the time delay between emitted light pulses and detected photons into time bins discretized over the whole clock cycle period for each pixel in the scene. Due to the presence of ambient sources, a proportion of the photons originate from either the signal (e.g. object or surface) or background sources (e.g. light emanating from the sun). The number of counts per time histogram bin provides information on the depth and reflectivity of a particular pixel in the field of view. The presence of a peak in the histogram typically indicates an object or surface is within the range of the lidar device, and, using the speed of light, one can convert the specific location of the time peak to a depth reading. If the material is semi-transparent, for example water, glass or camouflage, or the laser footprint is large then multiple peaks with different intensities may

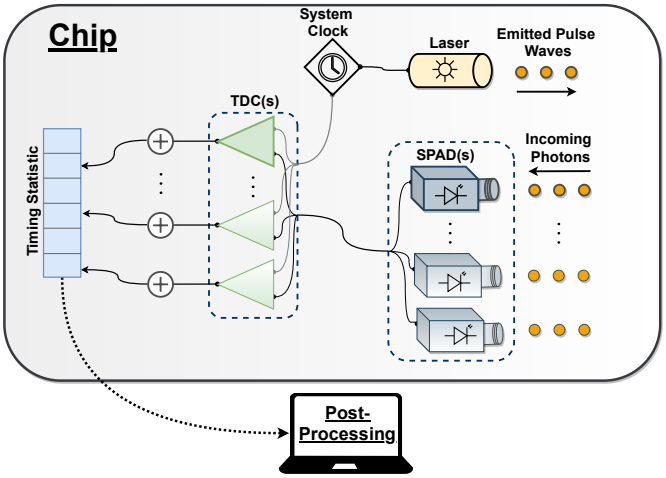

Figure 2.7: *A schematic of a typical lidar pixel where either one or multiple SPADs and TDCs are used.*

exist within a single pixel as demonstrated in Figure 2.6. If the number of photons detected are low, for example in a photon starved regime, then it is more efficient from a space complexity perspective to collect a timing statistic of each individual time-stamp for each pixel in the field of view.

Given a timing statistic, the depth image restoration task reduces down to inferring the positions and intensities of the peaks in the timing statistic for each pixel in the field of view. Let τ denote the physical time-stamp such that the discretized time-stamp is denoted $t = \frac{\tau}{\Delta\tau}$. Then, for an arbitrary pixel, the photon count at discretized time-stamp $t = 0, 1, \ldots, T-1$ can be modelled as a Poisson distribution [62, 63]:

$$y_{t_k}|(r, b, t_k) \sim \mathcal{P}(rh(t - t_k) + b), \tag{2.92}$$

where $r \geq 0$ denotes the reflectivity of the detected surface, $h(\cdot)$ is the impulse response of the system, b defines the level of background photons and t_k denotes the location of the kth surface in the pixel. The number of discretized time-stamp bins over the range of interest is denoted by T. For simplicity, here we assume that the integral of the impulse response $H = \sum_{t=0}^{T-1} h(t)$ is constant. If the lidar system is in free running mode where multiple acquisitions of a surface/object are obtained, then the interval $[0, 1, \ldots, T-1]$ can be thought of as circular in the sense that time-stamp T is equivalent to the time-stamp 0.

Alternatively, one can instead model the time of arrival of the pth photon detected for a single

pixel in the scene. We assume there are K distinct reflecting surfaces within the pixel, where α_k and α_0 denote the probability that the detected photon originated from the kth surface and background sources, respectively. Furthermore, it is assumed that for a single pixel, a total of N photons are detected during the whole acquisition window of the lidar device . Let $x_p \in \{0, 1, \ldots, T-1\}$ denote the time-stamp of the pth photon where $1 \leq p \leq N$, then x_p can be described by a mixture distribution [64]

$$\mathcal{P}(x_p|\alpha_0, ..., \alpha_K, t_1, ..., t_K) = \sum_{k=1}^{K} \alpha_k \mathcal{P}_s(x_p|t_k) + \alpha_0 \mathcal{P}_b(x_p), \tag{2.93}$$

where $\sum_{k=0}^{K} \alpha_k = 1$. The distribution of the photons originating from the signal and background are defined by the distribution $\mathcal{P}_s(x_p|t) = h(x_p - t)/H$ and the uniform distribution $\mathcal{P}_b(x_p) = 1/T$ over $[0, 1, \ldots, T-1]$, respectively. Often in practice, the signal distribution $\mathcal{P}_s$ is modelled either using a discretized Gaussian distribution over the interval $[0, 1, \ldots, T-1]$ or through the data driven impulse function which is calculated through experiments.

2.5.1 Data Transfer Bottleneck

The development of high rate, high resolution, low power ToF image sensors is challenging due to the large data volumes required. This causes a major data processing bottleneck on the device when either the number of photons per pixel N is large, the time resolution, $\Delta\tau$, is fine or the spatial resolution is high, as the space requirement, power consumption and computational burden of the depth reconstruction algorithms scale with these parameters [3].

Various existing methods have attempted to tackle the trade-off between depth resolution and computational/space complexities. A number of papers [65, 66, 67, 68, 69] propose methods to address the trade-off between depth resolution and the complexities associated with the TCSPC histogram. Henderson et al. [65] propose a method that employs a gated procedure to coarsely bin the detected photons, whilst Ren et al. [66] develop a sliding window approach to achieve high resolution depth. Walker et al. [67] calculate the depth directly from the photon time-stamps. However in all of these approaches, the approximations formed on-chip compromise the depth resolution of the image. Della Rocca et al. [68, 69] proposes to only collect the histograms of photon detections when there is a significant change of activity. This method reduces the data-transfer, as it is only required during specific moments in time. Similarly, Hutchings et al. [70] propose a method of discarding photon detections based on activity.

However, these methods can potentially remain idle when there is a small change in activity, and can also suffer from a loss of temporal resolution due to coarse histogram binning. Zhang et al. [71] propose a method of reducing the transfer of photon detections by performing a coarse to fine approximation of the ToF data. At each scale, a coarse histogram is constructed with a limited number of bins. Multiple histograms of increasing resolution have to be formed, hence the method has an increased total acquisition time and can also suffer from a loss of temporal resolution. In [72], Rapp et al. proposed a subtractive dithering for SPAD arrays that increases depth resolution without increasing the overall time-stamp resolution.

Compressive sensing strategies (see Section 2.3.1) have been successfully applied to lidar [73, 74], focusing on compressing the information across pixels. Kadambi et al. [73] propose to exploit the sparsity of natural scenes in some representation domain (e.g. wavelet transform) to reduce signal acquisition. The depth accuracy is limited by the level of amplitude noise and decay of the impulse response and is therefore limited to the case of one surface per pixel. Furthermore, the proposed method still requires large amounts of single-photon counting data to be transferred off-chip and therefore does not tackle the inherent data transfer bottleneck. In a similar vein, Halimi et al. [74] propose an adaptive sampling strategy that is scene dependent. By building up regions of interest and data driven depth maps in an iterative manner, they efficiently choose suitable scan positions to reduce acquisition time by up to 8 times in certain scenarios. However, the method relies on building TCSPC histograms and solving a maximization problem at each iteration of their adaptive algorithm. The method therefore has limitations for real-time processing especially when the amount of single-photon counting data is large. These compressive sensing based methods perform compression within the spatial domain and not throughout the depth domain and can therefore still suffer from data-transfer bottlenecks. Another approach to reduce the data transfer of the information needed to reconstruct the lidar image is to compress the data on-chip. As highlighted in [75], standard low-level data compression methods can be used to compress the data on-chip, however these methods can only offer up to a modest 50% data reduction and in some cases involve significant on-chip computation or there are limitations with respect to on-chip storage.

As discussed, the methods proposed to date do not tackle the data-transfer bottleneck without sacrificing depth resolution for compression and therefore the resulting depth reconstruction is sub-optimal. In Chapter 5, we will show that a compressive learning approach can vastly reduce the complexities associated with lidar imaging by constructing a sketch of the time-of-

flight data such that it retains all the information needed to efficiently infer the parameters θ of the lidar observation model. Fundamentally, the size of the sketch needs only to be of the order of the number of objects or surfaces in the scene. As lidar scenes typically consist of $K = 0, 1$ or 2 surfaces at most, the size of the compression attained by introducing a compressive learning approach to single photon counting lidar can be substantial as will be demonstrated.

Part I

Part One: New Models

Chapter 3
Compressive Independent Component Analysis

3.1 Introduction

The field of CL is still in its infancy with only a handful of developed schemes that cover compressive mixture models and compressive PCA. In this chapter, we look at the ICA model through the compressive learning lens and establish a compressive ICA framework that includes both practical algorithms and theoretical guarantees. The learning task of ICA was introduced in Section 2.1.6 where, in general, the distributional form of the independent components are left unspecified. To the same extent as PCA, the ICA model is typically defined in a semi-parametric manner. However, as is the case for PCA, a particular statistic of the data permits identifiability of the ICA parameters, i.e. the mixing matrix $\mathbf{M} \in \mathbb{R}^{d\times n}$, up-to scaling and permutation ambiguities. As discussed in Section 2.1.6, this particular statistic comes in the form of the kurtosis cumulant tensor $\mathscr{X} \in \mathfrak{C} \subset \mathbb{R}^{n\times n\times n\times n}$ associated with the data $\mathbf{X}$. As such, given that the number of independent components is denoted by n and the length of the data is denoted by the usual N, then the memory complexities of ICA typically scale with $\mathcal{O}\left(Nd + n^2\right)$ or $\mathcal{O}\left(n^4\right)$ depending on the particular class of ICA algorithm that is used. One can easily see that for large N or n, the memory demands of classical ICA methods quickly becomes infeasible. In this chapter, we show that the solutions to the cumulant based ICA model has particular structure and redundancies that induce a low-dimensional model set $\mathfrak{S}_\theta$ that resides in the space of cumulant tensors, i.e. $\mathfrak{S}_\theta \subset \mathfrak{C}$. To that end, we show that it is possible to compute a sketch of the data cumulants which has size $m \gtrsim n^2$ such that it encodes sufficient salient information to enable accurate inference of the ICA model parameters. Subsequently, the proposed compressive ICA scheme leads to orders of magnitude memory compression compared to existing ICA methods.

This chapter makes the following contributions:

- Focusing on the cumulant based ICA approach, we establish a low-dimensional model set

that resides in the larger cumulant space. We show that a sketch of size $m \gtrsim 2n(n+1)$, computed using sub-Gaussian measurements, satisfies a RIP on the model set $\mathfrak{S}_\theta$ with high probability. Furthermore, the RIP induces information preservation guarantees on the recovered cumulant tensor that shows the error between an arbitrary cumulant tensor and the recovered cumulant tensor is bounded linearly by modelling error and sampling noise. This establishes the existence of a robust decoder that, coupled with the sketching operator, forms a tractable compressive ICA scheme.

- In general, we do not have access to the true expected cumulant tensor but instead an approximation of the cumulant tensor formed by the finite samples. We establish an upper bound on the finite sampling error between the sketch of the expected cumulant tensor and the sketch of the approximated cumulant tensor. It is shown that the sampling error reduces as a function of the number of samples N.

- Two inherently different compressive ICA algorithms are proposed. The first algorithm is inspired by the greedy IPG schemes introduced in Section 2.3.1, where we design both a projection operator as well as a proxy projection operator that projects the updated tensor onto the model set $\mathfrak{S}_\theta$ at each iteration. The proxy projection operator is shown in practice to be robust to the non-convex landscape of the compressive ICA cost function. The second algorithm is an alternative steepest descent scheme that, in contrast to the IPG scheme, employs Riemannian optimization to optimize directly on the model set $\mathfrak{S}_\theta$.

- As part of the empirical results, we show that in practice a sharp phase transition, between successful and unsuccessful parameter estimation, occurs as the sketch size m grows. The region at which the transition transpires provides a pragmatic lower bound on the size of the sketch which one can use in practice. Furthermore, it is shown that this pragmatic lower bound coincides with the size of the sketch required to satisfy the RIP result. The loss of information incurred by taking a sketch of the ICA cumulants is demonstrated by comparing the statistical efficiency between the ICA estimates inferred by our compressive ICA algorithms and by existing algorithms that make use of the full data available.

This chapter is based of the work in [76] that appeared in the IEEE EUSIPCO conference 2019 and [77] that will appear at Information and Inference: A Journal of the IMA. The rest of this

chapter is organised as follows: In Section 3.2 the principles of compressive ICA are established were the low-dimensional model set is explicitly defined. In Section 3.3 we present and prove our main theoretical results of the chapter including the compressive ICA RIP, information preservation guarantees as well as finite sample effects. Two compressive ICA algorithms are proposed in Section 3.4 where the advantages and limitations are discussed. The empirical results including the phase transition, loss of information incurred by taking a sketch and experiments on real-world data are demonstrated in 3.5. Finally, we end the chapter in Section 3.6 with some concluding remarks.

3.2 Compressive Learning Principles for Cumulant ICA

In this section, the low-dimensional ICA model set $\mathfrak{S}_\theta$ that forms the basis of the compressive ICA framework is established. For convenience, let us recall some of the details of ICA that were introduced thoroughly in Section 2.1.6. Let $\mathbf{x} \in \mathbb{R}^d$ be a data point assumed to be zero mean, then the ICA model attempts to find a linear transformation $\mathbf{M} \in \mathbb{R}^{d\times n}$ such that

$$\mathbf{x} = \mathbf{M}\mathbf{s}, \tag{3.1}$$

where the individual entries or components of $\mathbf{s}$ are statistically independent, for example $\mathcal{P}(\mathbf{s}) = \prod_{i=1}^{n} \mathcal{P}_i(\mathbf{s}_i)$. As discussed in Section 2.1.6, prewhitening is a popular preprocessing trick. The process involves finding the matrix $\mathbf{V} \in \mathbb{R}^{n\times d}$ such that

$$\mathbf{z} = \mathbf{V}\mathbf{x} \;=\; \mathbf{V}\mathbf{M}\mathbf{s}, \tag{3.2}$$

where $\mathbf{z} \in \mathbb{R}^n$ has identity covariance matrix (uncorrelated). Prewhitening also handles the scenario when there is more mixtures than sources, i.e. $d > n$, by identifying the zero eigenvectors of the original data $\mathbf{x}$ (see Section 2.1.6). The matrix $\mathbf{Q} \coloneqq \mathbf{V}\mathbf{M}$ to be estimated is necessarily orthogonal . For the sake of presentation, we will subsequently consider the whitened version of the data for the remainder of this section and the corresponding whitened ICA equation

$$\mathbf{z} = \mathbf{Q}\mathbf{s}. \tag{3.3}$$

However, in Section 3.3.3, we propose 2 equivalent sketching frameworks that can incorporate both prewhitened and unwhitened data.

Cumulant or tensorial based ICA methods are of particular interest in this chapter. Given the model in Eqn 3.3 equating $\mathbf{z}$ to $\mathbf{s}$, then the following multilinear property holds for their associated 4th order cumulant tensors

$$\mathscr{Z} \;=\; \mathscr{S} \times_1 \mathbf{Q} \times_2 \mathbf{Q} \times_3 \mathbf{Q} \times_4 \mathbf{Q} \tag{3.4}$$

where $\times_j$ denotes the j-mode tensor-matrix product. The cumulant tensor $\mathscr{S}$ related to the independent components $\mathbf{s}$ is strictly diagonal due to the properties of cumulants defined in Eqn 2.27, i.e. $\mathscr{S}_{ijkl} = 0$ for all $ijkl \neq iiii$. Subsequently, we can state the cumulant based ICA model set:

$$\mathfrak{S}_\theta \;\coloneqq\; \{\mathscr{Z} \mid \mathscr{Z} \;=\; \mathscr{S} \times_1 \mathbf{Q} \times_2 \mathbf{Q} \times_3 \mathbf{Q} \times_4 \mathbf{Q},\, \mathscr{S} \in \mathfrak{D},\, \mathbf{Q} \in \mathrm{O}(n)\}\,, \tag{3.5}$$

where $\mathrm{O}(n)$ denotes the group of $n \times n$ orthogonal matrices and $\mathfrak{D} \subset \mathfrak{C}$ is the set of diagonal cumulant tensors defined by

$$\mathfrak{D} \;\coloneqq\; \{\mathscr{S} \mid \mathscr{S}_{ijkl} = 0\; \forall ijkl \neq iiii \text{ and } |\mathscr{S}_{iiii}| \geq \epsilon_{\mathscr{S}}\}\,. \tag{3.6}$$

Here, we have the additional requirement[1] that each diagonal cumulant is greater than or equal to a small constant $\epsilon_{\mathscr{S}} > 0$. Notably, any cumulant tensor $\mathscr{Z} \in \mathfrak{S}_\theta$ maximises any contrast function and hence minimises the associated theoretic measure and expected risk. The cumulant based ICA model set $\mathfrak{S}_\theta$ is itself a low-dimensional space residing in the space of cumulant tensors $\mathfrak{C}$. Specifically, $\mathfrak{S}_\theta$ can be described as the image of the product set of the set of $n \times n$ orthogonal matrices $\mathrm{O}(n)$ and the set of diagonal cumulant tensors $\mathfrak{D}$. We can therefore initially count the degrees of freedom of the model set $\mathfrak{S}_\theta$:

- $\mathfrak{D}$ - A maximum of n degrees of freedom on the leading diagonal.
- $O(n)$ - A maximum of $\frac{n(n-1)}{2}$ degrees of freedom [78].

In total, the model set has $\frac{n(n+1)}{2}$ degrees of freedom. In comparison, the space of 4th order cumulant tensors $\mathfrak{C}$, in which the model set resides, has $q \coloneqq \binom{n+3}{4} \approx \mathcal{O}(n^4)$ degrees of freedom. As the model set is of low complexity, in principle one could form a sketch of the

[1] A standard requirement in ICA is that at maximum one diagonal cumulant $\mathscr{S}_{iiii}$ can be zero which arises from the ICA assumption that at maximum one source signal s_i is Gaussian [13]. Here we have the slightly stronger assumption that all source signals are nongaussian and have nonzero kurtosis.

4th order cumulant tensor $\mathscr{Z}$ and estimate the parameters of the ICA model up to the usual ambiguities solely from the sketch. The sketch of the 4th order cumulant tensor $\mathscr{Z}$ is defined by

$$\mathbf{y}^{\mathbf{w}} = \mathcal{A}\left(\mathscr{Z}\right), \tag{3.7}$$

where $\mathbf{w}$ denotes that the sketch is acting on the whitened data $\mathbf{z}$. The computation of the sketch is very related to the sketching method of compressive PCA scheme discussed in Section 2.3.3. Akin to compressive PCA, the sketching operator $\mathcal{A}$ acts on the finite dimensional space of 4th order cumulant tensors instead of the infinite dimensional probability space which is left unspecified due to the semi-parametric nature of the ICA model. The ICA sketch defined in Eqn 3.7 draws strong connections to finite dimensional compressive sensing (see Section 2.3.1) where limited (random) measurements of a finite dimensional sparse vector are taken to reduce the sampling rate. As introduced in Eqn 2.43, the restricted isometry property (RIP) is a fundamental tool that is used to show that a sketching operator $\mathcal{A}$ stably embeds elements of the model set into a compressive domain $\mathbb{R}^m$, provided that the sketch dimension m is of sufficient size. In the case of compressive ICA, $\forall \mathscr{Z}_1, \mathscr{Z}_2 \in \mathfrak{S}_{\mathscr{H}}$ and an RIP constant $\delta \in (0, 1)$, then

$$(1-\delta)\|\mathscr{Z}_1 - \mathscr{Z}_2\|_F^2 \leq \|\mathcal{A}(\mathscr{Z}_1 - \mathscr{Z}_2)\|_2^2 \leq (1+\delta)\|\mathscr{Z}_1 - \mathscr{Z}_2\|_F^2 \tag{3.8}$$

provided that the sketch size m is of sufficient dimension and where $\|\cdot\|_F$ denotes the Frobenius norm. In many cases, the sketch size m is sufficient to be of the order of the degrees of freedom of the model set. In Section 2.3.1, it was discussed that if the lower RIP (LRIP) holds for a given sketching operator $\mathcal{A}$, e.g. the left side of Eqn 3.8, then there exists a robust decoder Δ that recovers a signal from the model set in a stable manner with respect to noise and signals that lie close to the model set. Moreover, it is proved in [38] that if the LRIP holds for the sketching operator $\mathcal{A}$ on the model set $\mathfrak{S}_\theta$ then the decoder Δ can be the constrained ℓ_2 optimization, for instance

$$\Delta\left(\mathbf{y}^{\mathbf{w}}, \mathcal{A}\right) \in \min_{\mathscr{Z} \in \mathfrak{S}_\theta} \|\mathbf{y}^{\mathbf{w}} - \mathcal{A}(\mathscr{Z})\|_2. \tag{3.9}$$

In principle, if the RIP can be proved for a sketching operator $\mathcal{A}$ on the ICA model set $\mathfrak{S}_\theta$, then we have an optimization strategy for solving the compressive ICA problem.

3.3 Compressive Independent Component Analysis Theory

We begin by explicitly defining the sketching operator $\mathcal{A} : \mathfrak{C} \mapsto \mathbb{R}^m$ as

$$\mathcal{A}(\mathscr{Z}) = \mathbf{A}\text{vec}(\mathscr{Z}), \tag{3.10}$$

where $\mathbf{A} \in \mathbb{R}^{m \times q}$ and vec denotes the vectorization operator. Here we assume $\mathbf{A}$ is some random measurement matrix where the entries $\mathbf{A}_{ij}$ are sampled according to some distributing law, $\mathbf{A}_{ij} \sim \Lambda$. In this chapter, we consider two randomized linear dimension reduction maps, namely the Gaussian map and the subsampled randomized Hadamard transform (SRHT) stated below. The compressive ICA (CICA) RIP, our main result stated in Theorem 1, is proved using the Gaussian map, however fast Johnson-Lindenstrauss transforms (FJLT), for instance the SRHT, still work in practice as will be discussed in Section 3.5.

3.3.0.1 Gaussian Maps

For convenience, we recall the traditional randomized linear dimension reduction map of Gaussian matrices. The Gaussian matrix $\mathbf{A} \in \mathbb{R}^{m \times q}$ has entries that follow

$$\mathbf{A}_{ij} \sim \mathcal{N}\left(0, m^{-\frac{1}{2}}\right). \tag{3.11}$$

Gaussian maps typically require $\mathcal{O}(mq)$ in memory as well as exhibiting a computational complexity of $\mathcal{O}\left(mq\right)$.

3.3.0.2 Subsampled Randomized Hadamard Transform

The SRHT is an instance of a FJLT that approximates the properties of the full Gaussian map [79]. Here $\mathbf{A} \in \mathbb{R}^{m \times q}$ is defined as

$$\mathbf{A} = \sqrt{\frac{q}{m}}\mathbf{RHD}, \tag{3.12}$$

where

- $\mathbf{D} \in \mathbb{R}^{q \times q}$ is a diagonal matrix whose elements are independent random signs $\{1, -1\}$;
- $\mathbf{H} \in \mathbb{R}^{q \times q}$ is a normalised Walsh-Hadamard matrix that is scaled by $p^{-\frac{1}{2}}$ so it is an

orthogonal matrix;

- $\mathbf{R} \in \mathbb{R}^{m \times q}$ is a matrix consisting of a subset of m randomly sampled rows from the $q \times q$ identity matrix.

The SRHT is particularly cheaper to compute and store in comparison to the Gaussian map. As we do not explicitly store $\mathbf{H}$, the SRHT only requires $\mathcal{O}(m+q)$ in memory [80]. In addition, the computational complexity of computing the sketch reduces to $\mathcal{O}(q\log(m))$ in comparison to using the Gaussian map [53, 80]. Below we state our main theoretical result of the chapter.

Theorem 1 (Compressive ICA RIP)**.** *Denote by $\mathcal{A}$ the Gaussian map sketching operator defined in Eqn 3.11. Then $\forall \mathscr{Z}_1, \mathscr{Z}_2 \in \mathfrak{S}_\theta$, the sketching operator $\mathcal{A}$ satisfies the RIP in Eqn 3.8 with constant $\delta \in (0,1)$ and probability $1-\xi$ provided that*

$$m \geq \frac{C}{\delta^2} \max\left\{2n(n+1)\log(C_0), \log\left(\frac{6}{\xi}\right)\right\}, \tag{3.13}$$

where $C = C(\xi) > 0$ is a constant dependent on ξ and $C_0 = C_0(\epsilon_{\mathscr{S}}) > 0$ is constant that is dependent on $\epsilon_{\mathscr{S}}$ defined in Lemma 2.

The proof of Theorem 1 is detailed in Section 3.3.1.

Corollary 1 (Information Preservation)**.** *Let $\mathscr{Z}^* \in \mathfrak{C}$ be an arbitrary 4th order cumulant tensor and denote $\mathbf{y}^{\boldsymbol{w}} = \mathcal{A}(\mathscr{Z}^*) + \mathbf{e}$ where $\mathbf{e} \in \mathbb{R}^m$ is some additive noise. Furthermore, let $\tilde{\mathscr{Z}} \coloneqq \Delta(\mathbf{y}^{\boldsymbol{w}}, \mathcal{A})$ denote the solution to Eqn 3.9. Given that $\mathcal{A}$ satisfies the RIP in Theorem 1, then with probability $1-\xi$*

$$\|\mathscr{Z}^* - \tilde{\mathscr{Z}}\|_F \leq \min_{\mathscr{Z} \in \mathfrak{S}_\theta} \left(2\|\mathscr{Z}^* - \mathscr{Z}\|_F + \frac{2}{\sqrt{1-\delta}}\|\mathbf{A}\,\mathrm{vec}\,(\mathscr{Z}^* - \mathscr{Z})\|_2\right) + \frac{2}{\sqrt{1-\delta}}\|\mathbf{e}\|_2 + \nu, \tag{3.14}$$

where $0 < \nu \leq 1$ is a small positive constant.

Proof. Given the LRIP in Theorem 1, we use Theorem 7 in [38] to obtain our result. □

The proof of Theorem 1 uses covering numbers and ϵ-nets of the normalized secant set of $\mathfrak{S}_\theta$ that were partially introduced in Section 2.3.1. For convenience, we recall the definitions below.

Definition 4 (Secant Set)**.** *The secant set of a set $\mathfrak{S}_\theta$ is defined as*

$$\mathfrak{S}_\theta - \mathfrak{S}_\theta \;\coloneqq\; \{\mathscr{Y} = \mathscr{Z}_1 - \mathscr{Z}_2 \mid \mathscr{Z}_1, \mathscr{Z}_2 \in \mathfrak{S}_\theta\}\,. \tag{3.15}$$

Definition 5 (Normalised Secant Set)**.** *The normalized secant set $\mathfrak{N}(\mathfrak{S}_\theta - \mathfrak{S}_\theta)$ of a set $\mathfrak{S}_\theta$ is defined as*

$$\mathfrak{N}(\mathfrak{S}_\theta - \mathfrak{S}_\theta) := \{\mathscr{Y}/\|\mathscr{Y}\|_F \mid \mathscr{Y} \in (\mathfrak{S}_\theta - \mathfrak{S}_\theta) \setminus \{\mathbf{0}\}\}\,, \tag{3.16}$$

where $\mathbf{0}$ defines the zero tensor.

Definition 6. *(Covering number) Let $\epsilon > 0$. The covering number, denoted CN$(\mathfrak{S}_\theta, \|\cdot\|, \epsilon)$, of a set $\mathfrak{S}_\theta$ is the minimum number of closed balls of radius ϵ, with respect to the norm $\|\cdot\|$, with centres in $\mathfrak{S}_\theta$ needed to cover $\mathfrak{S}_\theta$. The set of centres of these balls is a minimal ϵ-net for $\mathfrak{S}_\theta$.*

Lemma 2 (Covering number of $\mathfrak{N}(\mathfrak{S}_\theta - \mathfrak{S}_\theta)$)**.** *The covering number of $\mathfrak{N}(\mathfrak{S}_\theta - \mathfrak{S}_\theta)$ with respect to the Frobenius norm $\|\cdot\|_F$ is*

$$CN\left(\mathfrak{N}(\mathfrak{S}_\theta - \mathfrak{S}_\theta), \|\cdot\|_F, \epsilon\right) \leq \left(\frac{C_0}{\epsilon}\right)^{2n(n+1)}, \tag{3.17}$$

where $C_0 = C_0(\epsilon_{\mathscr{S}}) > 0$ is some constant.

Proof. See Appendix 3. □

Definition 7. *(Upper box counting dimension) The upper box counting dimension of a set S is defined as*

$$dim_B(S) := \limsup_{\epsilon \to 0} \log[CN(S, \|\cdot\|, \epsilon)]/\log[1/\epsilon]. \tag{3.18}$$

3.3.1 Proof of Theorem 1

Proof. To prove a RIP exists for the ICA model set $\mathfrak{S}_\theta$ using the sketching operator $\mathcal{A}$ defined in Eqn 3.10, we follow a similar line of argument to [35, 81] by using an ϵ-covering of $\mathfrak{N}(\mathfrak{S}_\theta - \mathfrak{S}_\theta)$ to extend the concentration results of the random Gaussian matrix $\mathbf{A}$ uniformally over the whole low-dimensional set. Specifically, we use the *Recipe* framework proposed by Puy *et al.* [37], to formulate the compressive ICA RIP proof. The proof is separated by showing that the following assumptions hold:

(A1) The normalised secant set, denoted $\mathfrak{N}(\mathfrak{S}_\theta - \mathfrak{S}_\theta)$, has finite upper-box counting dimension $\dim_{\mathrm{B}}(\mathfrak{N}(\mathfrak{S}_\theta - \mathfrak{S}_\theta))$ which is strictly bounded by $s \geq 1$, $\dim_{\mathrm{B}}(\mathfrak{N}(\mathfrak{S}_\theta - \mathfrak{S}_\theta)) < s$

(A2) The sketching operator $\mathcal{A}$ satisfies the concentration inequalities defined in Eqns 2.44 and 2.45 [37].

We begin with Assumption **(A1)**. Using Lemma 2 and the definition of the upper box counting dimension in Definition 7, it can be seen that $\dim_{\mathrm{B}}\left(\mathfrak{N}\left(\mathfrak{S}_\theta - \mathfrak{S}_\theta\right)\right) \leq 2n(n+1)$, so for any $s > 2n(n+1)$ Assumption **(A1)** is satisfied. To prove Assumption **(A2)**, we have the following definition.

Definition 8. *(Subgaussian random variable) A subgaussian random variable X is a random variable that satisfies*

$$(\mathbb{E}|X|^t)^{1/t} \leq C_1\sqrt{t} \text{ for all } t \geq 1, \tag{3.19}$$

with $C_1 > 0$. The subgaussian norm of X, denoted by $\|X\|_{\Psi_2}$ is the smallest C_1 for which the last property holds, i.e.,

$$\|X\|_{\Psi_2} := sup_{t\geq 1}\left\{t^{-1/2}(\mathbb{E}|X|^t)^{1/t}\right\}. \tag{3.20}$$

Let $\mathbf{A}_i$ denote the ith row of the random Gaussian matrix $\mathbf{A}$. Then we use the fact [82, 37] that

$$\|\mathbf{A}_i^T \operatorname{vec}(\mathscr{Z})\|_{\Psi_2} \leq D\|\mathscr{Z}\|_F \tag{3.21}$$

for all $\mathscr{Z} \in \mathfrak{C}$, where $D > 0$ is an absolute constant. Therefore Assumption **A2** is satisfied. Finally, using Theorem 8 of [37], we get the desired RIP result in Theorem 1. □

3.3.2 Finite Sample Effects

In practice, the sketch is constructed from a finite set of data $\{\mathbf{z}_i\}_{i=1}^N$ such that

$$\mathbf{y}_N^{\mathbf{w}} = \frac{1}{N}\sum_{i=1}^{N}\Phi^{\mathbf{w}}(\mathbf{z}_i), \tag{3.22}$$

where $\Phi^{\mathbf{w}}(\cdot)$ is the feature function discussed in Section 2.3 acting on the whitened data $\mathbf{z}$. For compressive ICA we can explicitly define the feature function, acting on the whitened data, as

$$\Phi_j^{\mathbf{w}}(\mathbf{z}) = \langle \mathbf{A}_j, \mathbf{z}^{\otimes^4}\rangle_F, \tag{3.23}$$

for $j = 1, \ldots, m$, where $\mathbf{A}_j \in \mathbb{R}^q$ are the rows of a Gaussian matrix $\mathbf{A}$ and $\langle \cdot \rangle$ denotes the Frobenius inner product. Furthermore, for shorthand we denote $\mathbf{z}^{\otimes^4} = \mathbf{z} \otimes \mathbf{z} \otimes \mathbf{z} \otimes \mathbf{z}$. In other words, the feature function is taking random quartics of the data point $\mathbf{z}$. Note that the empirical sketch $\mathbf{y}_N^{\mathbf{w}}$ is equivalent to $\mathbf{y}_N^{\mathbf{w}} = \mathcal{A}\left(\hat{\mathscr{Z}}\right)$, as specified in Eqn 2.63, where $\hat{\mathscr{Z}}$ is the finite data approximation of the 4th order cumulant tensor $\mathscr{Z}$ defined by

$$\begin{aligned}\hat{\mathscr{Z}}^4_{ijkl} = \frac{1}{N}\sum_{p=1}^{N} z_i^p z_j^p z_k^p z_l^p - \frac{1}{N^2}\sum_{p=1}^{N} z_i^p z_j^p \sum_{p=1}^{N} z_k^p z_l^p - \frac{1}{N^2}\sum_{p=1}^{N} z_i^p z_k^p \sum_{p=1}^{N} z_j^p z_l^p \\ - \frac{1}{N^2}\sum_{p=1}^{N} z_i^p z_l^p \sum_{p=1}^{N} z_j^p z_k^p,\end{aligned} \tag{3.24}$$

where z_i^p denotes the ith element of the pth finite sample. In this case, the error $\mathbf{e}$ defined in Theorem 1 can be attributed to the finite sample effects of approximating the true 4th order cumulant tensor $\mathscr{Z}$ from finite data. We now state our final result of this section.

Theorem 2 (Finite Sample Effects)**.** *Let $\mathcal{A}(\mathscr{Z}) = \mathbf{A}\,\mathrm{vec}(\mathscr{Z})$ denote the sketching operator where $\mathbf{A}_{ij} \sim \mathcal{N}\left(0, m^{-\frac{1}{2}}\right)$. Furthermore, let the independent components $\mathbf{s}$ have bounded support such that $\|\mathscr{S}\|_F \leq R$. Given that $\hat{\mathscr{Z}}$ is the finite approximation 4th order cumulant tensors computed from the random draw of finite samples $\mathbf{z}_1, \ldots, \mathbf{z}_N$, then with probability at least $1 - \rho - \xi$*

$$\|\mathcal{A}(\mathscr{Z}) - \mathcal{A}(\hat{\mathscr{Z}})\|_2 \leq \frac{CR\left(1 + \sqrt{2\log(1/\rho)}\right)}{\sqrt{N}}. \tag{3.25}$$

Proof. See Appendix 3.A.2. □

3.3.3 Discussion on Unwhitened Data

The results in this section are all based on proving a RIP on the model set $\mathfrak{S}_\theta$ defined in Eqn 3.5, where it is assumed the data $\mathbf{x}$ has been prewhitened to reduce the ICA model to $\mathbf{z} = \mathbf{Q}\mathbf{s}$ as discussed in Section 2.1.6. The prewhitening stage removes some of the degrees of freedom within the ICA inference task as it is necessary to estimate an orthogonal mixing matrix $\mathbf{Q}$. In some sketching cases, we may only see the data once, for example in the streaming context [57], and therefore prewhitening may not be possible. The fact that we are now estimating an arbitrary mixing matrix $\mathbf{M}$ instead of an orthogonal mixing matrix $\mathbf{Q}$ increases the degrees of freedom from $\frac{n(n+1)}{2}$ to $n(n+1)$. As a result, we must sketch the unwhitened moment tensor

$\mathscr{X}$ such that

$$\mathbf{y}^{\mathbf{u}} = \mathcal{A}(\mathscr{X}), \tag{3.26}$$

where $\mathcal{A}(\cdot) = \mathbf{A}\,\mathrm{vec}(\cdot)$ and $\mathbf{A} \in \mathbb{R}^{m \times q}$ is a random matrix as defined in Eqn 3.10. Here $\mathbf{u}$ denotes that the sketch is acting on the unwhitened data $\mathbf{x}$. In addition, the feature function $\Phi^{\mathbf{u}}(\cdot)$ for the unwhitened data can be defined as

$$\Phi^{\mathbf{u}}(\mathbf{x}) = \begin{bmatrix} \langle \mathbf{A}_j, \mathbf{x}^{\otimes^4} \rangle_F \\ \mathbf{x}^{\otimes^2} \end{bmatrix}, \tag{3.27}$$

for $j = 1, \ldots, m$, where $\mathbf{A}_j \in \mathbb{R}^q$ are the rows of the matrix $\mathbf{A}$. Note that the feature function for the unwhitened data now includes quadratic moments, as well as random quartic moments, that are needed to estimate the mixing matrix $\mathbf{M}$ which has extra degrees of freedom. One could further reduce the size of the unwhitened sketch by either computing random quadratic moments or simply removing the symmetries of the second order moments, however the reduction in complexity is minimal. Recall from Eqn 2.24 that the mixing matrix $\mathbf{M}$ has the following decomposition [18]

$$\mathbf{M} = \mathbf{V}^{-1}\mathbf{Q} \tag{3.28}$$

where $\mathbf{V} = \mathbf{\Pi}^{-\frac{1}{2}}\mathbf{P}^T$ is computed via the eigendecomposition of the covariance matrix $\mathbb{E}[\mathbf{x}\mathbf{x}^T]$ (see Section 2.1.5). In Section 3.4.2, we develop a CICA algorithm that specifically handles unwhitened data.

3.4 CICA Algorithms

In this section we propose two distinct compressive ICA algorithms to estimate the mixing matrix $\mathbf{M}$ for both the whitened and unwhitened case.

3.4.1 Iterative Projection Gradient

In Section 2.3.1 we introduced the Iterative projection gradient (IPG) descent algorithm. IPG is a popular optimization scheme which enforces low-dimensional structure e.g. sparsity, rank, etc, by projecting the object of interest onto the model set $\mathfrak{S}_\theta$ after each subsequent gradient step. For the case of compressive ICA, we seek an orthogonal projection on to the ICA model set $\mathfrak{S}_\theta$. Formally, we can define an orthogonal projection operator $\mathcal{P}_{\mathfrak{S}_\theta} : \mathfrak{C} \mapsto \mathfrak{S}_\theta$ of a 4th

order cumulant tensor $\mathscr{Z}^*$ as

$$\mathcal{P}_{\mathfrak{S}_\theta}\left(\mathscr{Z}^*\right) \in \underset{\mathscr{Z}\in\mathfrak{S}_\theta}{\arg\min}\|\mathscr{Z}^* - \mathscr{Z}\|_F. \tag{3.29}$$

In other words, $\mathcal{P}_{\mathfrak{S}_\theta}$ projects the object $\mathscr{Z}^* \in \mathfrak{C}$ onto the element in the model set that is of minimum distance w.r.t the Frobenius norm. In practice, it is often difficult to find a projection operator that is both orthogonal and tractable in terms of computation. In [83, 84], Cardoso showed that the ICA model set $\mathfrak{S}_\theta$ is a subset of $\mathfrak{R} \cap \mathfrak{L}$ where $\mathfrak{R}$ is the set of rank-n tensors defined as

$$\mathfrak{R} := \{\mathscr{Z} \in \mathfrak{R} \mid \text{rank}(\bar{\mathbf{Z}}) = n\}, \tag{3.30}$$

where $\bar{\mathbf{Z}} \in \mathbb{R}^{n^2\times n^2}$ is the matrix formed by rearranging the elements of the tensor $\mathscr{Z}$ into a $n^2 \times n^2$ Hermitian matrix and where rank defines the standard matrix rank [83]. Moreover, $\mathfrak{L}$ is the set of super-symmetric tensors defined by

$$\mathfrak{L} := \{\mathscr{Z} \in \mathfrak{L} \mid \mathscr{Z}_{\text{perm}(ijkl)} = \mathscr{Z}_{ijkl}\} \tag{3.31}$$

where perm defines all permutations of the index $ijkl$. In fact, Cardoso proved in [84] that locally the converse is true, for instance within some neighbourhood of $\mathscr{Z}$ the following holds:

$$\mathfrak{R} \cap \mathfrak{L} \subseteq \mathfrak{S}_\theta. \tag{3.32}$$

Therefore, within some neighbourhood of $\mathscr{Z}^*$, projecting onto the ICA model set $\mathfrak{S}_\theta$ is equivalent to projecting onto $\mathfrak{R} \cap \mathfrak{L}$. Cadzow proved in [85] that alternate projections onto $\mathfrak{R}$ and $\mathfrak{L}$ is guaranteed to converge onto the intersection[2] $\mathfrak{R} \cap \mathfrak{L}$. Fundamentally, the projections onto $\mathfrak{R}$ (rank-n approximation) and $\mathfrak{L}$ (averaging over permutations), denoted by $\mathcal{P}_\mathfrak{R}$ and $\mathcal{P}_\mathfrak{L}$ respectively, are both simple to compute and are orthogonal. Alternate orthogonal projections onto $\mathfrak{R}$ and $\mathfrak{L}$ ensures a stable projection onto $\mathfrak{R} \cap \mathfrak{L}$ [85] which results in a projection onto the ICA model set $\mathfrak{S}_\theta$. Formally, we define the projection $\mathcal{P}_{\mathfrak{S}_\theta}$ below in Algorithm 1. In practice, Algorithm 1 converges to below a small tolerance in very few iterations (~ 10 iterations). We can now state our full CICA IPG algorithm detailed in Algorithm 2. Here the step size μ_j is computed optimally to guarantee convergence [32, 33], $\mathcal{A}^*$ denotes the adjoint sketching operator and β is a fixed shrinking step size parameter. The practicalities and computational complexity

[2] In general, rank forcing destroys symmetry while symmetrization destroys the rank-n property, therefore alternate projections are needed until convergence.

of the algorithm will be deferred until Section 3.4.3.2.

Algorithm 1 $\mathcal{P}_{\mathfrak{S}_\theta}$: Projection onto ICA Model Set

Require: Cumulant tensor $\mathscr{Z}^* \in \mathfrak{C}$
  while Not Converged **do**
    Project onto $\mathfrak{R}$: $\mathscr{Z}^1 = \mathcal{P}_{\mathfrak{R}}(\mathscr{Z})$ (Matricize $\mathscr{Z}$ into a $n^2 \times n^2$ Hermitian matrix and take a rank-n approximation using truncated SVD)
    Project onto $\mathfrak{L}$: $\mathscr{Z}^2 = \mathcal{P}_{\mathfrak{L}}(\mathscr{Z}^1)$ (Average across all permutations of perm$(ijkl)$ for all indices $ijkl$)
  end while
  Output: $\mathscr{Z} \in \mathfrak{S}_\theta$

Algorithm 2 $\texttt{CICA}_{\texttt{IPG}}$: Iterative Projection Gradient Descent Compressive ICA

Require: Initialisation $\mathscr{Z}^0$, tolerance ϵ and shrinking parameter β.
  while $\|\mathbf{y}^{\mathbf{w}} - \mathcal{A}(\mathscr{Z}^j)\|_2^2 > \epsilon$ **do**
    Compute $\mu_j = \dfrac{\|\mathcal{A}^*(\mathbf{y}^{\mathbf{w}} - \mathcal{A}(\mathscr{Z}^j))\|_F^2}{\|\mathbf{y}^{\mathbf{w}} - \mathcal{A}(\mathscr{Z}^j)\|_2^2}$
    while $\|\mathbf{y}^{\mathbf{w}} - \mathcal{A}(\mathscr{Z}^{j+1})\|_2^2 > \|\mathbf{y}^{\mathbf{w}} - \mathcal{A}(\mathscr{Z}^j)\|_2^2$ **do**
      $\mu_j \leftarrow \beta\mu_j$
      $\mathscr{Z}^{j+\frac{1}{2}} \leftarrow \mathscr{Z}^j + \mu_j \mathcal{A}^*\left(\mathbf{y}^{\mathbf{w}} - \mathcal{A}\left(\mathscr{Z}^j\right)\right)$
      $\mathscr{Z}^{j+1} \leftarrow \mathcal{P}_{\mathfrak{S}_\theta}\left(\mathscr{Z}^{j+\frac{1}{2}}\right)$
    end while
  end while
  Output: $\mathscr{Z} \in \Delta\left(\mathbf{y}^{\mathbf{w}}, \mathcal{A}\right)$

3.4.2 Unwhitened IPG

It was discussed in Section 3.3.3 that it is often convenient, from an online processing point of view, to directly sketch the unwhitened data $\mathbf{x}$ and its associated cumulant tensor $\mathscr{X}$. Using the properties of the matrix-tensor product [15], it can be seen that

$$\mathbf{A}\,\mathrm{vec}\left(\mathscr{X}\right) = \mathbf{A}\bar{\mathbf{V}}^{-1}\,\mathrm{vec}\left(\mathscr{Z}\right), \tag{3.33}$$

where $\bar{\mathbf{V}} \coloneqq \mathbf{V} \otimes \mathbf{V} \otimes \mathbf{V} \otimes \mathbf{V} \in \mathbb{R}^{d^4 \times n^4}$. As defined in Eqn 3.27, the unwhitened feature function $\Phi^{\mathbf{u}}$ includes the second order moment of $\mathbf{x}$, namely $\mathbf{x}^{\otimes^2}$. The empirical sketch $\hat{\mathbf{y}}_N^{\mathbf{u}}$ therefore includes the sample covariance $\hat{\Sigma} \coloneqq \frac{1}{N}\sum_{i=1}^N \mathbf{x}_i^{\otimes^2}$, which can be used to estimate an approximation of $\mathbf{V}$, denoted $\hat{\mathbf{V}}$, by using the eigenvalue decomposition of $\hat{\Sigma}$ [19] at the beginning of Algorithm 2. By denoting $\hat{\bar{\mathbf{V}}} \coloneqq \hat{\mathbf{V}} \otimes \hat{\mathbf{V}} \otimes \hat{\mathbf{V}} \otimes \hat{\mathbf{V}}$, the gradient step in Algorithm

2 can be replaced by

$$\mathscr{Z}^{j+\frac{1}{2}} = \mathscr{Z}^{j} + \mu_j \mathbf{A}^T \big(\mathbf{y}^{\mathbf{u}} - \mathbf{A}\hat{\bar{\mathbf{V}}}^{-1} \operatorname{vec}\big(\mathscr{Z}^{j}\big)\big), \tag{3.34}$$

as well as the associated step size μ_j and stopping criteria. As a result, the CICA IPG algorithm proceeds as normal by employing the original projection $\mathcal{P}_{\mathfrak{S}_\theta}$. For the scenario when there are much more mixed signals than independent components, i.e $d \gg n$, the unwhitened IPG algorithm can become expensive to run due to the matrix multiplication $\mathbf{A}\bar{\mathbf{V}}^{-1}$.

3.4.3 Alternating Steepest Descent

The second proposed algorithm comes in the form of an alternating steepest descent (ASD) scheme that is inherently different from the IPG method previously discussed. To see why, it is insightful to rewrite Eqn 3.9 in terms of the elements of the product set $\mathfrak{D}$ and $\mathrm{O}(n)$:

$$\min_{\substack{\mathbf{Q}^T\mathbf{Q}=I \\ \mathscr{S}\in\mathfrak{D}}} F\left(\mathscr{S}, \mathbf{Q}\right) = \|\mathbf{y}^{\mathbf{w}} - \mathcal{A}(\mathscr{S} \times_1 \mathbf{Q} \times_2 \mathbf{Q} \times_3 \mathbf{Q} \times_4 \mathbf{Q})\|_2^2, \tag{3.35}$$

where we have used the multilinear property discussed in Eqn 2.28. As the optimization problem is now explicitly defined by the mixing matrix $\mathbf{Q}$ and a sparse diagonal tensor $\mathscr{S}$, it is sufficient to optimise with respect to these parameters in an alternating steepest descent scheme. This approach contrasts the IPG scheme, as once we initialise the mixing matrix $\mathbf{Q}$ and the diagonal cumulant tensor $\mathscr{S}$ appropriately, then we can optimise directly on the model set $\mathfrak{S}_\theta$. We can initially state the ASD steps:

1. $\mathscr{S}^* = \min_{\mathscr{S}\in\mathfrak{D}} F(\mathscr{S}, \mathbf{Q})$

2. $\mathbf{Q}^* = \min_{\mathbf{Q}^T\mathbf{Q}=I} F(\mathscr{S}^*, \mathbf{Q})$

Note that the diagonal cumulant tensor $\mathscr{S} \in \mathfrak{D}$ can be simply reformulated as an n sparse vector with known support, therefore one can perform element-wise differentiation on the n entries $\mathscr{S}_{iiii}$ for $i = 1 : n$. The second step requires more attention as we have the constraint $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$ (i.e. $\mathbf{Q} \in \mathrm{O}(n)$). The set of $n \times n$ orthogonal matrices is an instance of a Stiefel manifold [86], therefore F is minimized directly on the Stiefel manifold.

3.4.3.1 Stiefel Manifold Optimisation

Given a feasible matrix $\mathbf{Q}$ and the gradient $\nabla_{\mathbf{Q}} F = \left(\frac{\partial F(\mathscr{S},\mathbf{Q})}{\partial \mathbf{Q}_{ij}}\right)$, define a skew-symmetric matrix $\mathbf{B}$ as

$$\mathbf{B} = \nabla_{\mathbf{Q}} F \mathbf{Q}^T - \mathbf{Q}(\nabla_{\mathbf{Q}} F)^T. \tag{3.36}$$

The update on the Stiefel manifold is determined by the Crank-Nicholson scheme [87] denoted

$$Y(\tau) = \mathbf{Q} - \frac{\tau}{2}\mathbf{B}(\mathbf{Q} + Y(\tau)) \tag{3.37}$$

where $Y(\tau) = (I + \frac{\tau}{2}\mathbf{B})^{-1}(I - \frac{\tau}{2}\mathbf{B})\mathbf{Q}$. The matrix $(I + \frac{\tau}{2}\mathbf{B})^{-1}(I - \frac{\tau}{2}\mathbf{B})$ is referred to as the Cayley transform [86] of $\mathbf{B}$. The descent curve $Y(\tau)$ has the following useful features

- $Y(\tau)$ is smooth on τ
- $Y(0) = \mathbf{Q}$
- $Y(\tau)^T Y(\tau) = \mathbf{Q}^T\mathbf{Q}$ for all $\tau \in \mathbb{R}$.

As a result, we perform a steepest descent on $\mathbf{Q}$ with line search along the descent curve $Y(\tau)$ with respect to τ. For more details on optimisation methods constrained to the Stiefel manifold refer to [86]. We can now state our second proposed CICA algorithm in Algorithm 3.

Algorithm 3 $\texttt{CICA}_{\texttt{ASD}}$: Alternating Steepest Descent Compressive ICA

Require: Initialisation $\mathscr{Z}^0 = \mathscr{S}^0 \times_1 \mathbf{Q}^0 \times_2 \mathbf{Q}^0 \times_3 \mathbf{Q}^0 \times_4 \mathbf{Q}^0$, tolerance ϵ and step size μ.
  while $\|\mathbf{y}^{\mathbf{w}} - \mathcal{A}(\mathscr{Z}_j)\|_2^2 > \epsilon$ **do**
    $\mathscr{S}^{j+1} = \mathscr{S}^j + \mu \nabla_{\mathscr{S}} F\left(\mathscr{S}^j, \mathbf{Q}^j\right)$
    while Perform line search **do**
      $Y(\tau) = \mathbf{Q} - \frac{\tau}{2}\mathbf{B}\left(\mathbf{Q} + Y\left(\tau\right)\right)$
      $\mathbf{Q}^{t+1} \leftarrow Y\left(\tau^*\right)$
    end while
    $\mathscr{Z}^{j+1} \leftarrow \mathscr{S}^{j+1} \times_1 \mathbf{Q}^{j+1} \times_2 \mathbf{Q}^{j+1} \times_3 \mathbf{Q}^{j+1} \times_4 \mathbf{Q}^{j+1}$
  end while

3.4.3.2 Practicalities

We start by stating the computational complexity of each proposed CICA algorithm. Here we assume that a fast SRHT, as discussed in 3.3.0.2, is used to compute the sketch. For the IPG scheme, the symmetry projection $\mathcal{P}_{\mathfrak{L}}$ costs $\mathcal{O}(n^4)$ flops through averaging along all index

permutations. A rank-r approximation of a general matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$ costs $\mathcal{O}(r^2(n+m))$ flops [88], therefore the rank projection operator $\mathcal{P}_{\mathfrak{R}}$ costs a total of $\mathcal{O}(n^4)$ flops. The gradient step in Algorithm 2 costs a total of $\mathcal{O}(q\log(m))$ flops due to the use of the sketching operator $\mathcal{A}(\mathscr{Z}^j)$ at each iteration which results in the IPG algorithm therefore having a total cost of $\mathcal{O}(q\log(m)+n^4))$ flops. In the second proposed ASD algorithm, the gradient step in terms of the diagonal tensor in Algorithm 3, again has a cost of $\mathcal{O}(q\log(m))$ flops. The line search $Y(\tau)$ costs a total of $\mathcal{O}(n^3)$ flops [86] resulting in the ASD algorithm having a computational complexity of $\mathcal{O}(q\log(m)+n^3)$. Note that both proposed CICA algorithms have computational complexity that is independent of the length of the data N which can be extremely large for modern day applications.

As is the case for the general ICA problem, the compressive ICA optimisation problem is non-convex and both algorithms proposed may be prone to converging to local minima. As a result, we consider the option of possible restarts at random initialisations to obtain a good solution. We also consider a proxy projection operator that employs a Jacobi diagonalisation, popular in many ICA algorithms (see for example [19, 89]), followed by a hard thresholding operator that forces the off-diagonal elements $ijkl \neq iiii$ to zero. The Jacobi diagonalisation consists of maximising a contrast function with respect to the matrix $\mathbf{Q}$ via consecutive Givens rotations. Here we consider the contrast function [19]:

$$\varrho(\mathbf{Q}) = \sum_{i=1}^{n} \left(\mathscr{Z} \times_1 \mathbf{Q}^T \times_2 \mathbf{Q}^T \times_3 \mathbf{Q}^T \times_4 \mathbf{Q}^T \right)^2_{iiii}. \tag{3.38}$$

The resulting Jacobi scheme produces an approximately diagonalised 4^{th} order cumulant tensor and therefore we apply a hard thresholding operator by forcing the off-diagonals to zero. This procedure projects the updated cumulant tensor onto a point on the model set $\mathfrak{S}_\theta$ at each iteration and therefore acts as a proxy projector. We have observed in practice that this proxy projection operator is less sensitive to the non-convex landscape of the optimization problem, which could be explained by the robustness of Given's rotations [19], hence multiple restarts are rarely required. The proxy projection operator, which we denote by $\hat{\mathcal{P}}_{\mathfrak{S}_\theta}$, costs $\mathcal{O}(n^4)$ flops for the Given's rotation scheme to approximately diagonalise the cumulant tensor [19], and $\mathcal{O}(n^4-n)$ flops for the thresholding of the cross-cumulants. Therefore in total the proxy IPG algorithm has approximately the same computational complexity as our previous IPG algorithm. Furthermore, Given's rotations can be used to define a block coordinate descent scheme over $\mathrm{O}(n)$ to further reduce the computational complexity (see for instance [90]), however we

leave implementation of such a scheme for future work.

3.5 Empirical Results

3.5.1 Phase Transition

Phase transitions are an integral part of analysis that is used frequently in the compressive sensing literature to show a sharp change in the probability of successful reconstruction of the low-dimensional object as the sketch size m increases [91]. The location at which the phase transition occurs can provide a tight bound on the required sketch size needed given the number of independent components n and further consolidates the theoretical bound of the RIP derived in Section 3.3. To set up the phase transition experiment, we constructed the expected cumulant tensor $\mathscr{S}$ of n Laplacian sources and transformed the tensor with an orthogonal mixing matrix $\mathbf{M}$ using the multilinear property in Eqn 2.28, resulting in an expected cumulant tensor $\mathscr{Z}$. For each number of independent components n, 250 Monte Carlo simulations on the mixing matrix $\mathbf{M}$ were executed for increasing sketch size m between 2 and 700. A successful reconstruction was determined if the Amari error[3] [92] between the true mixing matrix $\mathbf{M}$ and the estimated mixing matrix $\hat{\mathbf{M}}$, defined by

$$d_{\mathrm{A}}(\mathbf{M},\hat{\mathbf{M}}) = \frac{1}{2n}\sum_{i=1}^{n}\left(\frac{\sum_{j=1}^{n}|b_{ij}|}{\max_j|b_{ij}|}-1\right)+\frac{1}{2n}\sum_{j=1}^{n}\left(\frac{\sum_{i=1}^{n}|b_{ij}|}{\max_i|b_{ij}|}-1\right), \tag{3.39}$$

was smaller than $d_{\mathrm{A}}(\mathbf{M},\hat{\mathbf{M}}) \leq 10^{-6}$, where $b_{ij} = (\mathbf{M}\hat{\mathbf{M}}^{-1})_{ij}$. The probability of successful reconstruction was given by the number of successful reconstructions within the 250 Monte-Carlo tests. We use the IPG version of the CICA algorithm for these results, although the ASD version provides nearly exactly the same results. It is insightful to begin by fixing the number of sources, here $n = 8$, to highlight the sharp transition as shown in Figure 3.1. We highlight some important bounds including the multiples of 2 and 4 times the dimension of the model set $\mathfrak{S}_\theta$, depicted by the orange lines. For comparison, the dimension of the space of cumulant tensors $\mathfrak{C}$, in other words the size of the cumulant tensor, is shown by the red line. The phase transition occurs in between 2 and 4 times the model set dimension indicating that choosing $m \geq 2n(n+1)$ would be sufficient in successfully inferring the mixing matrix with high

[3]The Amari error is used widely in the ICA literature as it is both scale and permutation invariant, which are the two inherent ambiguities of ICA inference.

probability.

Figure 3.2 generalises the single phase transition result for the number of independent components varying between $n = 2$ and $n = 10$. Once again, the important bounds of the model set dimension (green), 2 and 4 multiples of the model set dimension (orange) and the dimension of the space of cumulant tensors (red) are shown. Figure 3.2 explicitly shows that the phase transition empirically occurs within the location of $m = n(n+1)$ and $m = 2n(n+1)$ and provides us with a tight practical lower bound of $m \geq 2n(n+1)$ on the sketch size for successful inference of the mixing matrix with high probability. Recall that in Theorem 1, the RIP holds when $m \gtrsim 2n(n+1)$. The location of the phase transition in the empirical results therefore further consolidates the theoretical result. For a given number of independent components n, the ratio between the upper orange line (4 times the model set dimension) and the red line (space of cumulant tensor dimension) provides a realistic compression rate in comparison to using the whole cumulant tensor of which many ICA techniques use. Importantly, as the number of independent components increases the ratio between these two lines decreases, resulting in further compression.

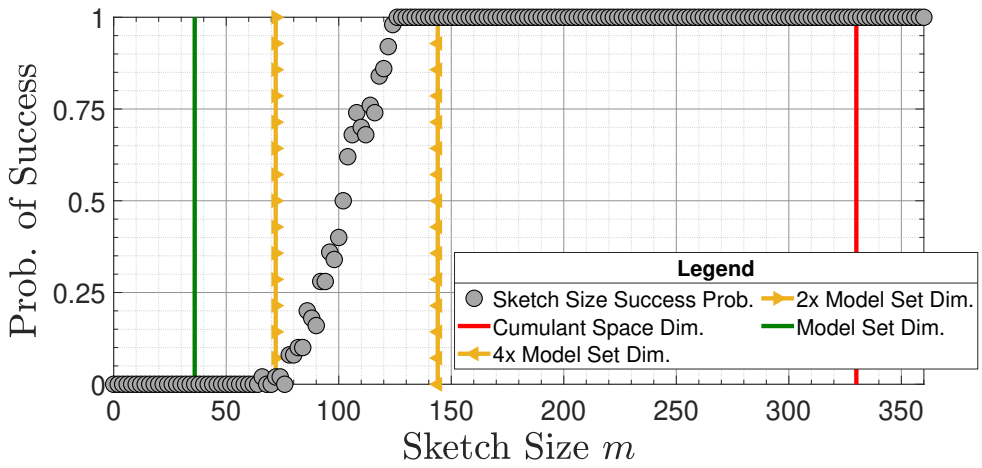

Figure 3.1: *A phase transition between unsuccessful and successful mixing matrix inference as the sketch size m increases and the number of independent components is fixed at $n = 8$.*

3.5.2 Statistical Efficiency

As was shown in Section 3.5.1, the potential compression rates of sketching the cumulant tensor are high which can lead to a significantly reduced memory requirement. In this section we numerically analyse the trade-off between the sketch size and the loss of information. The

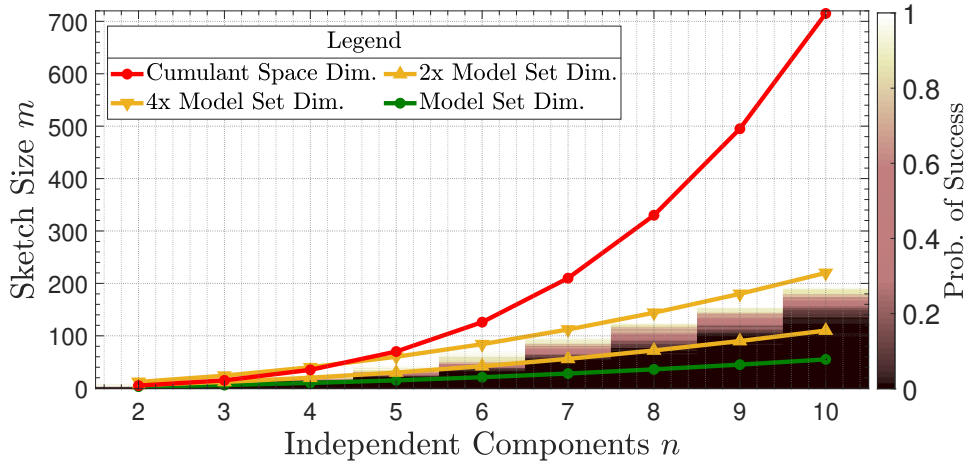

Figure 3.2: *A phase transition between unsuccessful and successful mixing matrix inference as the sketch size m and the number of independent components n increases.*

statistical efficiency of an estimator is a measure of the variability or quality of an unbiased estimator [20]. In Section 2.2.3 we introduced the the Cramér-Rao bound that provides a lower bound on the variability of an estimator and gives a best case scenario. For fair comparison, here we instead use the variability of an estimator inferred by an algorithm that explicitly makes use of the cumulant information to estimate the information loss of our proposed CICA algorithm. As such, we use Comon's ICA algorithm, detailed in [19], that minimizes a kurtosis based contrast function using a sequence of Given's rotations on pairwise cumulants as the approximate full data bound (e.g. no compression). We could have equivalently used the well-known Joint Approximation Diagonalization of Eigen-matrices (JADE) algorithm [89] or any other cumulant based ICA algorithm as the approximate bound, which gives similar results. To this end, we make use of the relative efficiency, defined as

$$e(\mathbf{M}_1, \mathbf{M}_2) = \frac{\text{var}\left(d_{\text{A}}(\mathbf{M}_\theta, \mathbf{M}_1)\right)}{\text{var}\left(d_{\text{A}}(\mathbf{M}_\theta, \mathbf{M}_2)\right)}, \tag{3.40}$$

where $d_{\text{A}}(\cdot, \cdot)$ is the Amari Error defined in Eqn 3.39 and $\mathbf{M}_\theta$ is the true mixing matrix. Denoting $\mathbf{M}_{\text{FD}}$ and $\mathbf{M}_{\text{CICA}}$ as the mixing matrix estimates of Comon's ICA algorithm (full data) and the proposed CICA algorithm, respectively, we expect $0 \leq e\left(\mathbf{M}_{\text{FD}}, \mathbf{M}_{\text{CICA}}\right) \leq 1$ as the Comon algorithm exhibits no compression and makes use of the full cumulant tensor available. As the relative efficiency $e\left(\mathbf{M}_{\text{FD}}, \mathbf{M}_{\text{CICA}}\right)$ approaches 1, the sketch estimate becomes more statistically efficient. We perform our efficiency test on $n = 6$ independent components of signal length $N = 1000$. The signal length does not affect the results as the dependence of N

drops out of the relative efficiency measure, for example see [76].

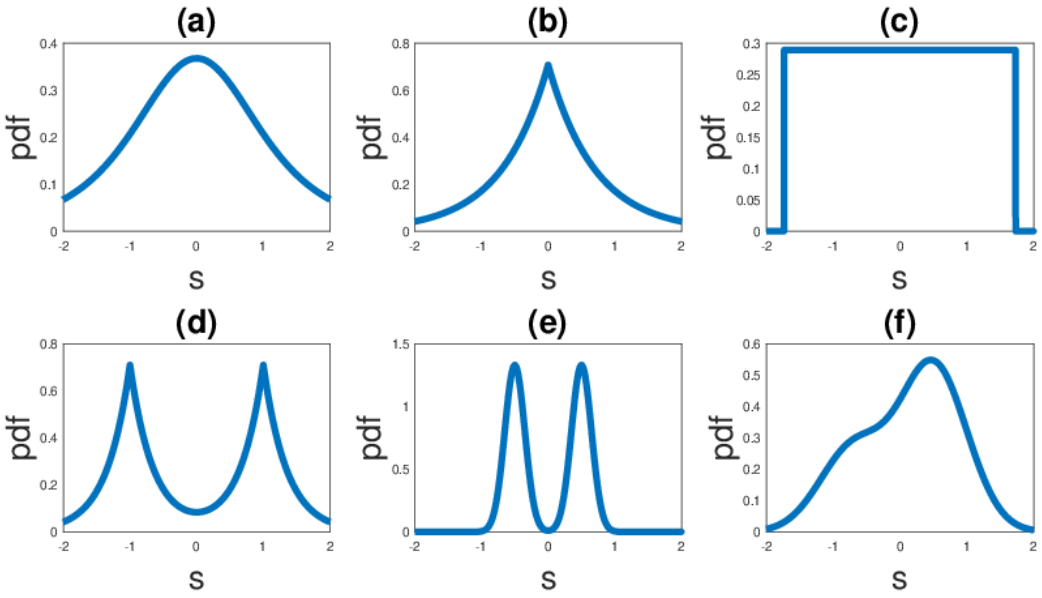

Figure 3.3: *(a) Student's t distribution* $(\nu = 3)$ *(b) Laplace distribution* $(\mu = 0, b = 1)$ *(c) continuous uniform distribution* $(a = -\sqrt{3}, b = \sqrt{3})$ *(d) mixture of 2 Laplaces* $(\mu_1, \mu_2 = -1, 1\ b_1 = b_2 = 1)$ *(e) symmetric bimodal mixture of Gaussians* $(\mu_1, \mu_2 = -1, 1\ \sigma_1 = \sigma_2 = 0.15)$ *(f) asymmetric unimodal mixture of Gaussians* $(\mu_1, \mu_2 = -0.7, 0.5\ \sigma_1 = \sigma_2 = 0.5)$

For each of the 100 Monte-Carlo simulations, the $n = 6$ independent components are randomly sampled [93] from a range of distributions with unique characteristics that are shown in Figure 3.3. The true mixing matrix $\mathbf{M}_\theta$ was sampled once and fixed throughout. For each sketch size m, 100 simulations were executed where the mixing matrix was estimated and the Amari error was calculated. The variance of the Amari errors was compared with the full data counterpart and plotted as the relative efficiency in Figure 3.4. Figure 3.4 shows the relative efficiency as the sketch size m increases. As m increases the relative efficiency approaches 1 (i.e. as statistically efficient as using the full cumulant tensor with no compression). It is evident that there is a trade-off between the rate of compression and the statistical efficiency, for instance, the smaller the sketch size the greater the loss of statistical efficiency. This is to be expected as the harsher you compress the data the more loss of information you experience. Nonetheless, the tradeoff is controlled, for example, a sketch of size $m = 100$ has a drop of around 40% of efficiency.

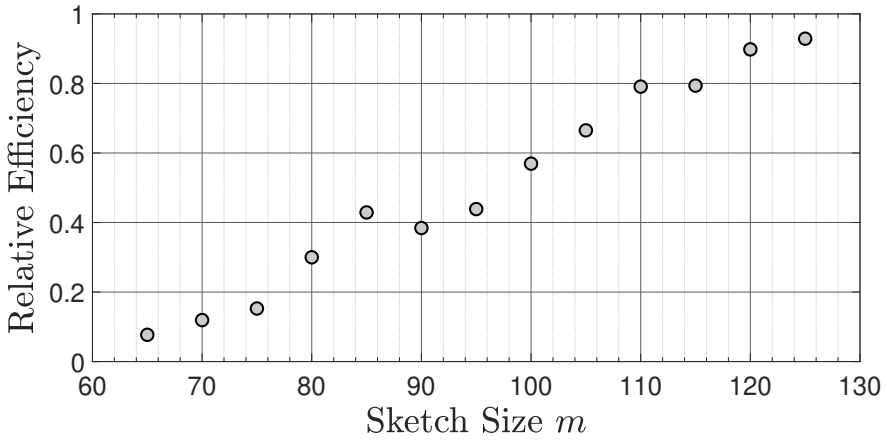

Figure 3.4: *The relative efficiency of the full data cumulant tensor (Comon's ICA) and sketch mixing matrix estimates for increasing sketch size m.*

3.5.3 Cylinder Velocity Field

We next analyse and compare the proposed CICA scheme on a dataset consisting of a flow field around a cylinder obstruction as depicted in Figure 3.5. Using ICA, one can obtain a model that describes the fluctuations of the streamwise velocity field around its mean value as a function of time. Details of the experimental set up can be seen in [94, 95]. The dataset is of size $\mathbf{X} \in \mathbb{R}^{100\times 14400}$ consisting of 14400 spatial locations over 100 time intervals. Here we compare our proposed CICA scheme with the well-known fast ICA algorithm [96], as well the JADE [89] and Comon algorithm [19] which, like the proposed CICA scheme, are cumulant based. An initial prewhitening stage inferred the prewhiten matrix $\mathbf{V} \in \mathbb{R}^{8\times 14400}$. Each algorithm then estimated the $\mathbf{Q} \in \mathbb{R}^{8\times 8}$, resulting in a mixing matrix estimate $\mathbf{M} = \mathbf{V}^{-1}\mathbf{Q}$. For the proposed CICA scheme, the IPG version was used with a SRHT matrix $\mathbf{A}$, however ASD version produces similar reconstructions. Figure 3.6 shows the 8 independent components which describe the fluctuations of the streamwise velocity around the cylinder obtained by Fast ICA, JADE, Comon and CICA, respectively. For our proposed CICA algorithm, a sketch of size $m = 114$ is used. Visually comparing the reconstructions, one can see that the CICA algorithm performs competitively with negligible artifacts present. In addition, the CICA scheme achieves a compression rate of approximately 3 in comparison to the other cumulant based ICA methods discussed.

Next, we compare the effect of the sketch size on the resulting reconstructions. A sketch size of $m = 72, 108$ and 144 are considered with the reconstructions shown in Figure 3.7. For

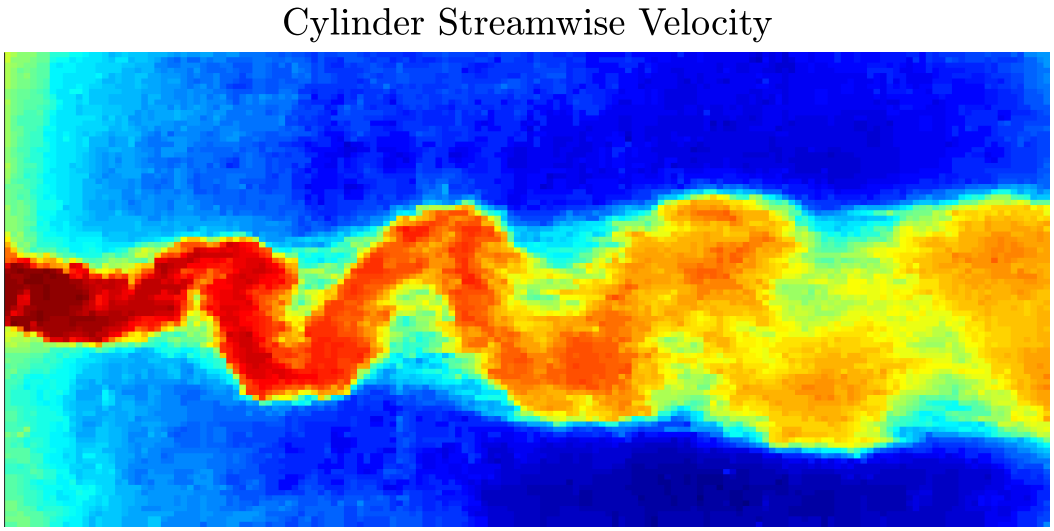

Figure 3.5: *The figure shows the velocity field around a cylinder for a fixed point in time.*

$m = 108$, the sketch is of sufficient size to successfully identify the unique fluctuations of the velocity field, however, due to the harsher compression rate some notable artefacts are present. For example, in the first and third fluctuations there are some oscillating type artifacts which can be attributed to the higher frequencies in the system. Furthermore, the sketch of size $m = 72$ fails to identity the main fluctuations of the velocity field.

3.6 Concluding Remarks

In this chapter, a low-dimensional model set was shown to exist for the cumulant based ICA problem. It was demonstrated theoretically that a RIP exists for the ICA model using Gaussian maps provided the sketch size was set proportionally to the model set dimensions, which in turn induced the existence of an instance optimal decoder. The theoretical results were empirically validated by showing the location of a sharp phase transition between a state of unsuccessful inference to a state of successful inference of the ICA mixing matrix as the sketch size increased. By considering optimising on the ambient cumulant tensor space or directly on the low-dimensional model set, we proposed two inherently different CICA algorithms. Using both synthetic and real data, we analysed the robustness of the proposed CICA algorithms and highlighted the effect of choosing the sketch size m. Furthermore, the particular branch of compressive learning was discussed that consists of sketching distribution free models (e.g. PCA, ICA) that leverage some intermediary statistic space, here the space of cumulant tensors,

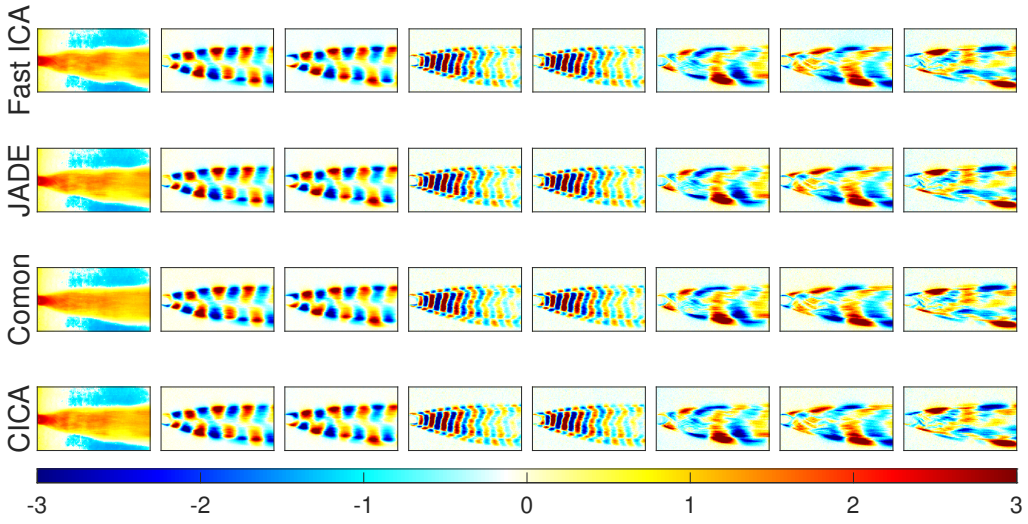

Figure 3.6: *From left to right the dominant fluctuations of the streamwise velocity field. From top to bottom the Fast ICA, JADE, Comon and CICA reconstructions.*

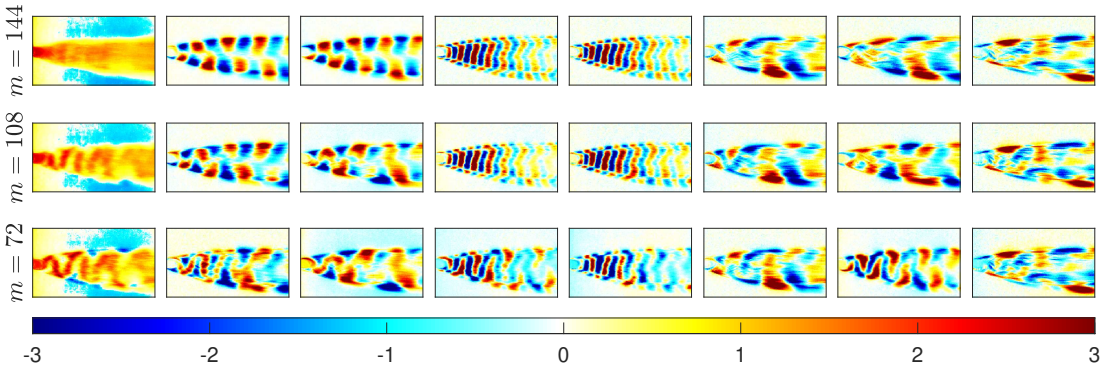

Figure 3.7: *The figure shows the effect of the sketch size on the reconstruction of the fluctuations. From top to bottom a sketch size of* $m = 144, 108$ *and* 72.

to form the sketch. This poses some interesting open questions that we attempt to answer in Chapter 4 on how to design a sketch given other distribution free models and how the low-dimension nature of the model set manifests itself structurally, in terms of sparsity, low rank, etc. to construct a practical sketching decoder.

3.A Appendix

3.A.1 Proof of Lemma 2

To prove Lemma 2, we use a similar line of argument to Clarkson in [97] by splitting the normalized secant set into the set of short and long secants parametrized by a distance η. First we state an important lemma on covering the model set intersected with the unit sphere in $\mathbb{R}^{\bar{n}}$, where $\bar{n} = n^4$, denoted by $\bar{\bar{\mathfrak{S}}}_\theta := \mathfrak{S}_\theta \cap \mathbb{S}^{\bar{n}-1}$ (e.g. $\|\mathscr{Z}\|_F = 1$), that will used later in the proof.

Lemma 3 (Covering number of $\bar{\bar{\mathfrak{S}}}_\theta$)**.** *The covering number of $\bar{\bar{\mathfrak{S}}}_\theta$ with respect to the Frobenius norm $\|\cdot\|_F$ is*

$$CN\left(\bar{\bar{\mathfrak{S}}}_\theta, \|\cdot\|_F, \epsilon\right) \leq \left(\frac{6}{\epsilon}\right)^{n(n+1)} \tag{3.41}$$

Proof. Recall that $\mathscr{Z} \in \bar{\bar{\mathfrak{S}}}_\theta$ has the decomposition $\mathscr{Z} = \mathscr{S} \times_1 \mathbf{Q} \times_2 \mathbf{Q} \times_3 \mathbf{Q} \times_4 \mathbf{Q}$ such that $\|\mathscr{Z}\|_F = 1$ where $\mathscr{S} \in \mathfrak{D}$ and $\mathbf{Q} \in \mathrm{O}(n)$. As the Frobenius norm is rotationally invariant then the following holds $\|\mathscr{Z}\|_F = \|\mathscr{S}\|_F = 1$ for all $\mathscr{Z} \in \bar{\bar{\mathfrak{S}}}_\theta$. Our argument constructs an ϵ-net for $\bar{\bar{\mathfrak{S}}}_\theta$ by covering the sets $\mathfrak{D}$ and $\mathrm{O}(n)$ respectively. As $\|\mathscr{Z}\|_F = 1 \implies \|\mathscr{S}\|_F = 1$, it is sufficient to consider $\bar{\bar{\mathfrak{D}}} := \mathfrak{D} \cap \mathbb{S}^{n-1}$. Then we take $\underline{\bar{\bar{\mathfrak{D}}}}$ to be an $\epsilon/2$- net for $\bar{\bar{\mathfrak{D}}}$. As $\bar{\bar{\mathfrak{D}}}$ is a n dimensional subspace, then

$$\mathrm{CN}\left(\bar{\bar{\mathfrak{D}}}, \|\cdot\|_F, \epsilon/2\right) \leq \left(\frac{6}{\epsilon}\right)^n.$$

Next, we cover the set of $n \times n$ orthogonal matrices denoted $\mathrm{O}(n)$. We follow a similar argument to [81, 35] by letting $\mathrm{Q}(n) := \{\mathbf{Y} \in \mathbb{R}^{n \times n} : \|\mathbf{Y}\|_{1,2} \leq 1\}$, where

$$\|\mathbf{Y}\|_{1,2} = \max_i \|\mathbf{Y}(:, i)\|_2 \tag{3.42}$$

is the maximum column norm of a matrix $\mathbf{Y}$. It is straightforward to see that $\mathrm{O}(n) \subset \mathrm{Q}(n)$ since the columns of an orthogonal matrix are unit normed. It can be seen in [35] that an $\epsilon/2$-net $\mathrm{O}(n)$, denoted by $\underline{\mathrm{O}}(n)$, has a covering number

$$\mathrm{CN}(\mathrm{O}(n), \|\cdot\|_{1,2}, \epsilon/2) \leq \left(\frac{6}{\epsilon}\right)^{n^2}.$$

Now let $\underline{\bar{\bar{\mathfrak{S}}}}_\theta := \{\underline{\mathscr{S}} \times_1 \underline{\mathbf{Q}} \times_2 \underline{\mathbf{Q}} \times_3 \underline{\mathbf{Q}} \times_4 \underline{\mathbf{Q}} : \underline{\mathscr{S}} \in \underline{\bar{\bar{\mathfrak{D}}}}, \underline{\mathbf{Q}} \in \underline{\mathrm{O}}(n)\}$, and remark that

$$\begin{aligned}\mathrm{CN}\big(\bar{\bar{\mathfrak{S}}}_\theta, \|\cdot\|_F, \epsilon\big) &\leq \mathrm{CN}\big(\bar{\bar{\mathfrak{D}}}, \|\cdot\|_F, \epsilon/2\big)\ \ \mathrm{CN}\big(\mathrm{O}(n), \|\cdot\|_{1,2}, \epsilon/2\big) \\ &\leq \Big(\frac{6}{\epsilon}\Big)^{n(n+1)}.\end{aligned}$$

It remains to show that for all $\mathscr{Z} \in \bar{\bar{\mathfrak{S}}}_\theta$ there exists $\underline{\mathscr{Z}} \in \underline{\bar{\bar{\mathfrak{S}}}}_\theta$ such that $\|\mathscr{Z} - \underline{\mathscr{Z}}\|_F \leq \epsilon$.

Fix $\mathscr{Z} \in \bar{\bar{\mathfrak{S}}}_\theta$ and note the decomposition $\mathscr{Z} = \mathscr{S} \times_1 \mathbf{Q} \times_2 \mathbf{Q} \times_3 \mathbf{Q} \times_4 \mathbf{Q}$. Then there exists $\underline{\mathscr{Z}} = \underline{\mathscr{S}} \times_1 \underline{\mathbf{Q}} \times_2 \underline{\mathbf{Q}} \times_3 \underline{\mathbf{Q}} \times_4 \underline{\mathbf{Q}} \in \underline{\bar{\bar{\mathfrak{S}}}}_\theta$ with $\underline{\mathscr{S}} \in \underline{\bar{\bar{\mathfrak{D}}}}$ and $\underline{\mathbf{Q}} \in \underline{O}(n)$ obeying $\|\mathscr{S} - \underline{\mathscr{S}}\|_F \leq \epsilon/2$ and $\|\mathbf{Q} - \underline{\mathbf{Q}}\|_{1,2} \leq \epsilon/2$. This gives

$$\begin{aligned}\|\mathscr{Z} - \underline{\mathscr{Z}}\|_F &= \|\mathscr{S} \times_1 \mathbf{Q} \times_2 \mathbf{Q} \times_3 \mathbf{Q} \times_4 \mathbf{Q} - \underline{\mathscr{S}} \times_1 \underline{\mathbf{Q}} \times_2 \underline{\mathbf{Q}} \times_3 \underline{\mathbf{Q}} \times_4 \underline{\mathbf{Q}}\|_F \\ &= \|\mathscr{S} \times_1 \mathbf{Q} \times_2 \mathbf{Q} \times_3 \mathbf{Q} \times_4 \mathbf{Q} + (\mathscr{S} \times_1 \underline{\mathbf{Q}} \times_2 \underline{\mathbf{Q}} \times_3 \underline{\mathbf{Q}} \times_4 \underline{\mathbf{Q}} \\ &\quad - \mathscr{S} \times_1 \underline{\mathbf{Q}} \times_2 \underline{\mathbf{Q}} \times_3 \underline{\mathbf{Q}} \times_4 \underline{\mathbf{Q}}) - \underline{\mathscr{S}} \times_1 \underline{\mathbf{Q}} \times_2 \underline{\mathbf{Q}} \times_3 \underline{\mathbf{Q}} \times_4 \underline{\mathbf{Q}}\|_F \\ &= \|\mathscr{S} \times_1 (\mathbf{Q} - \underline{\mathbf{Q}}) \times_2 (\mathbf{Q} - \underline{\mathbf{Q}}) \times_3 (\mathbf{Q} - \underline{\mathbf{Q}}) \times_4 (\mathbf{Q} - \underline{\mathbf{Q}}) \\ &\quad + (\mathscr{S} - \underline{\mathscr{S}}) \times_1 \underline{\mathbf{Q}} \times_2 \underline{\mathbf{Q}} \times_3 \underline{\mathbf{Q}} \times_4 \underline{\mathbf{Q}}\|_F \\ &\leq \|\mathscr{S} \times_1 (\mathbf{Q} - \underline{\mathbf{Q}}) \times_2 (\mathbf{Q} - \underline{\mathbf{Q}}) \times_3 (\mathbf{Q} - \underline{\mathbf{Q}}) \times_4 (\mathbf{Q} - \underline{\mathbf{Q}})\|_F \\ &\quad + \|(\mathscr{S} - \underline{\mathscr{S}}) \times_1 \underline{\mathbf{Q}} \times_2 \underline{\mathbf{Q}} \times_3 \underline{\mathbf{Q}} \times_4 \underline{\mathbf{Q}}\|_F\end{aligned}$$

The first part of the last line gives

$$\begin{aligned}\|\mathscr{S} \times_1 \cdots \times_4 (\mathbf{Q} - \underline{\mathbf{Q}})\|_F &= \|\mathrm{vec}\,(\mathscr{S} \times_1 (\mathbf{Q} - \underline{\mathbf{Q}}) \times_2 (\mathbf{Q} - \underline{\mathbf{Q}}) \times_3 (\mathbf{Q} - \underline{\mathbf{Q}}) \times_4 (\mathbf{Q} - \underline{\mathbf{Q}}))\|_2 \\ &= \|(\mathbf{Q} - \underline{\mathbf{Q}}) \otimes (\mathbf{Q} - \underline{\mathbf{Q}}) \otimes (\mathbf{Q} - \underline{\mathbf{Q}}) \otimes (\mathbf{Q} - \underline{\mathbf{Q}})\mathrm{vec}(\mathscr{S})\|_2 \\ &\leq \|(\mathbf{Q} - \underline{\mathbf{Q}}) \otimes (\mathbf{Q} - \underline{\mathbf{Q}}) \otimes (\mathbf{Q} - \underline{\mathbf{Q}}) \otimes (\mathbf{Q} - \underline{\mathbf{Q}})\|_2 \|\mathscr{S}\|_F \\ &= \|(\mathbf{Q} - \underline{\mathbf{Q}})\|_2^4 \\ &\leq \|(\mathbf{Q} - \underline{\mathbf{Q}})\|_{1,2}^4 \\ &\leq (\epsilon/2)^4 \\ &\leq \epsilon/2\end{aligned}$$

From line 1 to 2, the identity on pages [477-478] of [98] was used. From line 2 to 3 we have used the Cauchy-Schwarz inequality, from line 3 to 4 we have used the equality $\|\mathbf{A} \otimes \mathbf{B}\| = \|\mathbf{A}\|\|\mathbf{B}\|$

and from line 4 to 5 we have used the identity in [81]. Finally, notice that as $\mathbf{Q}$ is orthogonal

$$\|(\mathscr{S} - \underline{\mathscr{S}}) \times_1 \underline{\mathbf{Q}} \times_2 \underline{\mathbf{Q}} \times_3 \underline{\mathbf{Q}} \times_4 \underline{\mathbf{Q}}\|_F = \|(\mathscr{S} - \underline{\mathscr{S}})\|_F = \epsilon/2. \tag{3.43}$$

Therefore

$$\|\mathscr{Z} - \underline{\mathscr{Z}}\|_F \leq \epsilon/2 + \epsilon/2 = \epsilon \tag{3.44}$$

□

Continuing, we let $\Omega \coloneqq O(n) \times \mathfrak{D}$ define the product set between the set of $n \times n$ orthogonal matrices $O(n)$ and the set of super symmetric cumulant tensors defined in Eqn 2.29 and define the map $f : \Omega \mapsto \mathfrak{S}_\theta$ by

$$f(u) = \mathscr{S} \times_1 \mathbf{Q} \times_2 \mathbf{Q} \times_3 \mathbf{Q} \times_4 \mathbf{Q}, \tag{3.45}$$

for all $u \coloneqq (\mathbf{Q}, \mathscr{S}) \in \Omega$. Let $\mathscr{Z} = f(u)$ be the tensor corresponding to the image of the map f. It is insightful to decompose the normalised secant set $\mathfrak{N}(\mathfrak{S}_\theta - \mathfrak{S}_\theta)$ into the set of long and short secants parametrised by some distance η [97]. The set of long secants of $\mathfrak{S}_\theta$ is defined as

$$\mathfrak{N}_\eta(\mathfrak{S}_\theta - \mathfrak{S}_\theta) \coloneqq \left\{ \frac{\mathscr{Z}_1 - \mathscr{Z}_2}{\|\mathscr{Z}_1 - \mathscr{Z}_2\|_F} \;\middle|\; \mathscr{Z}_1, \mathscr{Z}_2 \in \mathfrak{S}_\theta,\ \|\mathscr{Z}_1 - \mathscr{Z}_2\|_F > \eta \right\}. \tag{3.46}$$

Furthermore, the set of short secants $\mathfrak{N}^c_\eta(\mathfrak{S}_\theta - \mathfrak{S}_\theta) = \mathfrak{N}(\mathfrak{S}_\theta - \mathfrak{S}_\theta) \setminus \mathfrak{N}_\eta(\mathfrak{S}_\theta - \mathfrak{S}_\theta)$ is the complement to the set of long secants defined by

$$\mathfrak{N}^c_\eta(\mathfrak{S}_\theta - \mathfrak{S}_\theta) \coloneqq \left\{ \frac{\mathscr{Z}_1 - \mathscr{Z}_2}{\|\mathscr{Z}_1 - \mathscr{Z}_2\|_F} \;\middle|\; \mathscr{Z}_1 \neq \mathscr{Z}_2 \in \mathfrak{S}_\theta,\ \|\mathscr{Z}_1 - \mathscr{Z}_2\|_F \leq \eta \right\}. \tag{3.47}$$

Remark 4. *As the model set $\mathfrak{S}_\theta$ is conic (see Lemma 9), it is sufficient to cover the normalised secant set of $\bar{\mathfrak{S}}_\theta \coloneqq \mathfrak{S}_\theta \cap \mathfrak{B}_1(0)$, where $\mathfrak{B}_1(0)$ denotes the unit Frobenius ball centred at 0, since we have $\mathfrak{N}(\mathfrak{S}_\theta - \mathfrak{S}_\theta) = \mathfrak{N}\left(\bar{\mathfrak{S}}_\theta - \bar{\mathfrak{S}}_\theta\right)$.*

As a result we can decompose the normalised secant set as follows

$$\begin{aligned} \mathfrak{N}(\mathfrak{S}_\theta - \mathfrak{S}_\theta) &= \mathfrak{N}\left(\bar{\mathfrak{S}}_\theta - \bar{\mathfrak{S}}_\theta\right) \\ &= \mathfrak{N}_\eta\left(\bar{\mathfrak{S}}_\theta - \bar{\mathfrak{S}}_\theta\right) \cup \mathfrak{N}^c_\eta\left(\bar{\mathfrak{S}}_\theta - \bar{\mathfrak{S}}_\theta\right) \\ &\subseteq \mathfrak{N}_\eta\left(\bar{\mathfrak{S}}_\theta - \bar{\mathfrak{S}}_\theta\right) \cup \mathfrak{N}^c_\eta\left(\bar{\mathfrak{S}}_\theta - \bar{\bar{\mathfrak{S}}}_\theta\right), \end{aligned} \tag{3.48}$$

We begin by covering the set of long secants $\mathfrak{N}_\eta\left(\bar{\mathfrak{S}}_\theta - \bar{\mathfrak{S}}_\theta\right)$.

Lemma 4 (Long Secants Covering Number)**.** *Let $\underline{\bar{\mathfrak{S}}}_\theta$ be an $\epsilon\gamma$-cover for $\bar{\mathfrak{S}}_\theta$. Then $\mathfrak{N}\left(\underline{\bar{\mathfrak{S}}}_\theta - \underline{\bar{\mathfrak{S}}}_\theta\right)$ is an ϵ-cover for $\mathfrak{N}_{4\gamma}\left(\bar{\mathfrak{S}}_\theta - \bar{\mathfrak{S}}_\theta\right)$ with associated covering number of*

$$CN\left(\mathfrak{N}_{4\gamma}\left(\bar{\mathfrak{S}}_\theta - \bar{\mathfrak{S}}_\theta\right), \|\cdot\|_F \epsilon\gamma\right) \leq \left(\frac{6}{\epsilon\gamma}\right)^{2n(n+1)}. \tag{3.49}$$

Proof. Lemma 4.1 in [97] states that if $\underline{\bar{\mathfrak{S}}}_\theta$ is a generalised $\epsilon\gamma$- cover of $\bar{\mathfrak{S}}_\theta$, then $\mathfrak{N}\left(\underline{\bar{\mathfrak{S}}}_\theta - \underline{\bar{\mathfrak{S}}}_\theta\right)$ is a generalised ϵ-cover for $\mathfrak{N}_{4\gamma}\left(\bar{\mathfrak{S}}_\theta - \bar{\mathfrak{S}}_\theta\right)$. Using the covering number of $\bar{\bar{\mathfrak{S}}}_\theta$ from Lemma 3 we get the result. □

Continuing, we cover the set of short secants. We begin by stating some preliminary lemmas.

Lemma 5 (Taylor Approximation Error)**.** *Let $f : \Omega \mapsto \mathfrak{S}_\theta$ be defined as in Eqn 3.45 and let Df_u define the first order differential of f evaluated at the point u. Further assume that $\|\mathscr{S}\|_F \leq R$. Then $\forall u, u' \in \Omega$, $\|u - u'\| \leq 2\epsilon_0$, we have*

$$\left\|f(u) - f(u') - Df_{u'}^T(u - u')\right\|_F \leq C_1 \left\|u - u'\right\|_2^2, \tag{3.50}$$

where $C_1 = n^2(n+1)^2 \max\{3R, 1\}$

Proof. w.l.o.g consider the vectorized function $\tilde{f}(u) \coloneqq \mathrm{vec}(f(u))$ such that

$$\begin{aligned}\tilde{f}(u) =& \text{vec}\left(\mathscr{S} \times_1 \mathbf{Q} \times_2 \mathbf{Q} \times_3 \mathbf{Q} \times_4 \mathbf{Q}\right)\\ =& \mathbf{Q} \otimes \mathbf{Q} \otimes \mathbf{Q} \otimes \mathbf{Q}\ \text{vec}\left(\mathscr{S}\right).\end{aligned}$$

Using Taylor's theorem [99, p. 110] of $\tilde{f}$ evaluated at the point $u' \in \Omega$, we get

$$\left\|\tilde{f}(u) - \tilde{f}(u') - D\tilde{f}_{u'}^T(u - u')\right\|_2 \leq \frac{1}{2}\left\|(u - u')^T H\tilde{f}_\xi (u - u')\right\|_2. \tag{3.51}$$

where $D\tilde{f}_u$ and $H\tilde{f}_u$ denote the Jacobian and Hessian of $\tilde{f}$ evaluated at u and $\xi = \lambda u + (1 - \lambda)u' \in \Omega$, for $\lambda \in (0, 1)$, denotes a point on the line segment between u and u'. For shorthand

let $h = u - u'$, and denote the integer $T \coloneqq \frac{n(n+1)}{2}$, we then have

$$\begin{aligned}
\left\| h^T H \tilde{f}_\xi h \right\|_2 &= \left\| \sum_{i=1}^{T} \sum_{j=1}^{T} h_i h_j \frac{\partial^2 \tilde{f}}{\partial u_i \partial u_j}(\xi) \right\|_2 \\
&\leq T^2 \max_{i,j} \left\| h_i h_j \frac{\partial^2 \tilde{f}}{\partial u_i \partial u_j}(\xi) \right\|_2 \\
&\leq T^2 \left(\max_i |h_i| \right)^2 \max_{i,j} \left\| \frac{\partial^2 \tilde{f}}{\partial u_i \partial u_j}(\xi) \right\|_2 \\
&= T^2 \|h\|_\infty^2 \max_{i,j} \left\| \frac{\partial^2 \tilde{f}}{\partial u_i \partial u_j}(\xi) \right\|_2 \\
&\leq T^2 \|h\|_2^2 \max_{i,j} \left\| \frac{\partial^2 \tilde{f}}{\partial u_i \partial u_j}(\xi) \right\|_2 ,
\end{aligned}$$

where $h_i = (u_i - u_i')$. w.l.o.g let $\xi = (\mathbf{Q}, \mathscr{S})$, we have that

$$\max_{i,j} \left\| \frac{\partial^2 \tilde{f}}{\partial u_i \partial u_j}(\xi) \right\|_2 = \max \left\{ \overset{①}{\max_{i,j,k,\ell}} \left\| \frac{\partial^2 \tilde{f}}{\partial \mathbf{Q}_{ij} \partial \mathbf{Q}_{kl}}(\xi) \right\|_2 , \overset{②}{\max_{i,j,k}} \left\| \frac{\partial^2 \tilde{f}}{\partial \mathbf{Q}_{ij} \partial \mathscr{S}_{kkkk}}(\xi) \right\|_2 , \max_{i,j} \overset{③}{\left\| \frac{\partial^2 \tilde{f}}{\partial \mathscr{S}_{iiii} \partial \mathscr{S}_{jjjj}}(\xi) \right\|_2} \right\} \tag{3.52}$$

① It can be seen that (see Section 3.A.5)

$$\frac{\partial^2 \tilde{f}}{\partial \mathbf{Q}_{ij} \partial \mathbf{Q}_{k\ell}}(\xi) = \Pi_{ijk\ell} \text{vec}\left(\mathscr{S}\right), \tag{3.53}$$

where

$$\begin{aligned}
\Pi_{ijk\ell} = \; & \mathbf{E}^{ij} \otimes \mathbf{E}^{k\ell} \otimes \mathbf{Q} \otimes \mathbf{Q} \; + \; \mathbf{E}^{ij} \otimes \mathbf{Q} \otimes \mathbf{E}^{k\ell} \otimes \mathbf{Q} \; + \; \mathbf{E}^{ij} \otimes \mathbf{Q} \otimes \mathbf{Q} \otimes \mathbf{E}^{k\ell} \\
& + \; \mathbf{E}^{k\ell} \otimes \mathbf{E}^{ij} \otimes \mathbf{Q} \otimes \mathbf{Q} \; + \; \mathbf{Q} \otimes \mathbf{E}^{ij} \otimes \mathbf{E}^{k\ell} \otimes \mathbf{Q} \; + \; \mathbf{Q} \otimes \mathbf{E}^{ij} \otimes \mathbf{Q} \otimes \mathbf{E}^{k\ell} \\
& + \; \mathbf{E}^{k\ell} \otimes \mathbf{Q} \otimes \mathbf{E}^{ij} \otimes \mathbf{Q} \; + \; \mathbf{Q} \otimes \mathbf{E}^{k\ell} \otimes \mathbf{E}^{ij} \otimes \mathbf{Q} \; + \; \mathbf{Q} \otimes \mathbf{Q} \otimes \mathbf{E}^{ij} \otimes \mathbf{E}^{k\ell} \\
& + \; \mathbf{E}^{k\ell} \otimes \mathbf{Q} \otimes \mathbf{Q} \otimes \mathbf{E}^{ij} \; + \; \mathbf{Q} \otimes \mathbf{E}^{k\ell} \otimes \mathbf{Q} \otimes \mathbf{E}^{ij} \; + \; \mathbf{Q} \otimes \mathbf{Q} \otimes \mathbf{E}^{k\ell} \otimes \mathbf{E}^{ij}
\end{aligned}$$

and the matrix $\mathbf{E}^{ij} = \mathbf{e}_i \mathbf{e}_j^T$, where $\mathbf{e}_i$ is the ith unit basis vector. Using the properties of

the Kronecker product and the triangle inequality we get

$$\begin{aligned}\left\|\frac{\partial^2 \tilde{f}}{\partial \mathbf{Q}_{ij}\partial \mathbf{Q}_{kl}}(\xi)\right\|_2 &\leq 12\left\|\mathbf{E}^{ij}\right\|_2\left\|\mathbf{E}^{kl}\right\|_2\|\mathbf{Q}\|_2^2\|\mathscr{S}\|_F \\ &= 12\|\mathscr{S}\|_F.\end{aligned}$$

Assuming that the diagonal tensor has bounded support $\|\mathscr{S}\|_2 \leq R$, then it follows that

$$\max_{i,j,k,\ell}\left\|\frac{\partial^2 \tilde{f}}{\partial \mathbf{Q}_{ij}\partial \mathbf{Q}_{kl}}(\xi)\right\|_2 \leq 12R. \tag{3.54}$$

② It can be seen that (see Section 3.A.5)

$$\frac{\partial^2 \tilde{f}}{\partial \mathbf{Q}_{ij}\partial \mathscr{S}_{kkkk}}(\xi) = \Gamma_{ij}\mathbf{e}_k, \tag{3.55}$$

where and

$$\begin{aligned}\Gamma_{ij} = \mathbf{E}^{ij}\otimes\mathbf{Q}\otimes\mathbf{Q}\otimes\mathbf{Q} \;+\; \mathbf{Q}\otimes\mathbf{E}^{ij}\otimes\mathbf{Q}\otimes\mathbf{Q} \\ +\; \mathbf{Q}\otimes\mathbf{Q}\otimes\mathbf{E}^{ij}\otimes\mathbf{Q} \;+\; \mathbf{Q}\otimes\mathbf{Q}\otimes\mathbf{Q}\otimes\mathbf{E}^{ij}.\end{aligned}$$

Similarly to ①, we get

$$\max_{i,j,k}\left\|\frac{\partial^2 \tilde{f}}{\partial \mathbf{Q}_{ij}\partial \mathscr{S}_{kkkk}}(\xi)\right\|_2 \leq 4 \tag{3.56}$$

③ It can be easily shown that

$$\frac{\partial^2 \tilde{f}}{\partial \mathscr{S}_{iiii}\partial \mathscr{S}_{jjjj}}(\xi) = \mathbf{0}, \tag{3.57}$$

therefore

$$\max_{i,j}\left\|\frac{\partial^2 \tilde{f}}{\partial \mathscr{S}_{iiii}\partial \mathscr{S}_{jjjj}}(\xi)\right\|_2 = 0. \tag{3.58}$$

It therefore follows that

$$\max_{i,j}\left\|\frac{\partial^2 \tilde{f}}{\partial u_i \partial u_j}(\xi)\right\|_2 = \max\left\{12R, 4\right\}, \tag{3.59}$$

and,

$$\left\|\tilde{f}(u) - \tilde{f}(u') - D\tilde{f}_{u'}^{T}(u-u')\right\|_2 \leq n^2(n+1)^2 \max\{3R, 1\} \left\|u-u'\right\|_2^2. \tag{3.60}$$

□

Lemma 6 (Bounded Curvature)**.** *Let $f : \Omega \mapsto \mathfrak{S}_\theta$ be defined as in Eqn 3.45 and let Df_u define the first order differential of f evaluated at the point u. Further assume that $\|\mathscr{S}\|_F \leq R$. Then $\forall u, u' \in \Omega$, $\|u-u'\| \leq 2\epsilon_0$, we have*

$$\|Df_u - Df_{u'}\|_F \leq C_2 \left\|u-u'\right\|_2, \tag{3.61}$$

where $C_2 = 2C_1$

Proof. Using the mean value theorem [99], it can be shown that,

$$\left\|D\tilde{f}_u \;-\; D\tilde{f}_{u'}\right\|_2 \leq \left\|H\tilde{f}_{\xi}^{T}(u-u')\right\|_2 \tag{3.62}$$

for some $\xi = \lambda u + (1-\lambda)u' \in \Omega$, for $\lambda \in (0,1)$. Then using the same argument as in the proof of Lemma 5, it can easily shown that

$$\left\|D\tilde{f}_u \;-\; D\tilde{f}_{u'}\right\|_2 \leq 2C_1 \left\|u-u'\right\|_2, \tag{3.63}$$

giving $C_2 = 2C_1$. □

Lemma 7 (Bounded Gradient)**.** *Let $f : \Omega \mapsto \mathfrak{S}_\theta$ be defined as in Eqn 3.45 and let Df_u define the first order differential of f evaluated at the point u. Further assume, as in Eqn 2.29, that $|\mathscr{S}_{iiii}| \geq \epsilon_{\mathscr{S}} (> 0)$ $\forall i$. Then $\forall u \in \Omega$*

$$\left\|Df_u^{\dagger}\right\|_F \leq C_3, \tag{3.64}$$

where $C_3 = 2\epsilon_{\mathscr{S}}$

Proof. Similar to Lemma 5, we consider the vectorized function $\tilde{f}(u) \coloneqq \mathrm{vec}\,(f(u))$ w.l.o.g. It can be seen that the 1st order differential (see Section 3.A.4) has the following decomposition

$$D\tilde{f}(u) = \left[\frac{\partial \tilde{f}}{\partial \mathbf{Q}}(u), \frac{\partial \tilde{f}}{\partial \mathscr{S}}(u)\right], \tag{3.65}$$

where

$$\frac{\partial \tilde{f}}{\partial \mathbf{Q}_{ij}}(u) \;=\; \Gamma_{ij} \operatorname{vec}(\mathscr{S}). \tag{3.66}$$

Furthermore, the partial derivative with respect to the super symmetric cumulant tensor $\mathscr{S}$ is defined as

$$\frac{\partial \tilde{f}}{\partial \mathscr{S}}(u) \;=\; \bar{\mathbf{Q}}.$$

where $\bar{\mathbf{Q}} \coloneqq \mathbf{Q} \otimes \mathbf{Q} \otimes \mathbf{Q} \otimes \mathbf{Q}$. Equivalently, Eqn 3.64 can be rewritten as

$$\min_{\|\Delta u\|=1} \left\| D\tilde{f}(u)^T \Delta u \right\|_2 \geq C_3, \tag{3.67}$$

where $\Delta u = (\Delta \mathbf{Q}, \Delta \mathscr{S})$. We therefore have

$$\begin{aligned}
\left\| D\tilde{f}(u)^T \Delta u \right\|_2^2 &= \left\| \frac{\partial \tilde{f}}{\partial \mathbf{Q}}(u)^T \Delta \mathbf{Q} \right\|_F^2 + \left\| \frac{\partial \tilde{f}}{\partial \mathscr{S}}(u)^T \Delta \mathscr{S} \right\|_2^2 \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \left\| \frac{\partial \tilde{f}}{\partial \mathbf{Q}_{ij}}(u)^T \Delta \mathbf{Q}_{ij} \right\|_F^2 + \left\| \frac{\partial \tilde{f}}{\partial \mathscr{S}}(u)^T \Delta \mathscr{S} \right\|_2^2 \\
&= (\star).
\end{aligned}$$

As f is equivariant in $\mathbf{Q}$, we can set $\mathbf{Q} = \mathbf{I}_n$ w.l.o.g. As a result $\bar{\mathbf{Q}} = \mathbf{I}$ and Γ_{ij} reduces to

$$\begin{aligned}
\Gamma_{ij} \;=\; & \mathbf{E}^{ij} \otimes \mathbf{I}_n \otimes \mathbf{I}_n \otimes \mathbf{I}_n \;+\; \mathbf{I}_n \otimes \mathbf{E}^{ij} \otimes \mathbf{I}_n \otimes \mathbf{I}_n \\
& + \mathbf{I}_n \otimes \mathbf{I}_n \otimes \mathbf{E}^{ij} \otimes \mathbf{I}_n \;+\; \mathbf{I}_n \otimes \mathbf{I}_n \otimes \mathbf{I}_n \otimes \mathbf{E}^{ij}.
\end{aligned}$$

For shorthand, let $\operatorname{vec}(\mathcal{T}) = \Gamma_{ab} \operatorname{vec}(\mathscr{S})$ and noting that $\mathbf{E}^{ab} = \mathbf{e}_a \mathbf{e}_b^T$, we have

$$\begin{aligned}\mathcal{T}_{ijk\ell} &= \sum_{p=1}^{n} \left(\mathbf{E}^{ab}_{ip}\mathbf{I}_{jp}\mathbf{I}_{kp}\mathbf{I}_{\ell p} + \mathbf{I}_{ip}\mathbf{E}^{ab}_{jp}\mathbf{I}_{kp}\mathbf{I}_{\ell p} + \mathbf{I}_{ip}\mathbf{I}_{jp}\mathbf{E}^{ab}_{kp}\mathbf{I}_{\ell p} + \mathbf{I}_{ip}\mathbf{I}_{jp}\mathbf{I}_{kp}\mathbf{E}^{ab}_{\ell p} \right) \mathscr{S}_{pppp} \\ &= \sum_{p=1}^{n} \left(\delta_{ai}\delta_{bp}\delta_{jp}\delta_{kp}\delta_{\ell p} + \delta_{ip}\delta_{aj}\delta_{bp}\delta_{kp}\delta_{\ell p} + \delta_{ip}\delta_{jp}\delta_{ak}\delta_{bp}\delta_{\ell p} + \delta_{ip}\delta_{jp}\delta_{kp}\delta_{a\ell}\delta_{bp} \right) \mathscr{S}_{pppp} \\ &= \sum_{p=1}^{n} \left(\delta_{ai}\delta_{jp}\delta_{kp}\delta_{\ell p} + \delta_{ip}\delta_{aj}\delta_{kp}\delta_{\ell p} + \delta_{ip}\delta_{jp}\delta_{ak}\delta_{\ell p} + \delta_{ip}\delta_{jp}\delta_{kp}\delta_{a\ell} \right) \delta_{bp}\mathscr{S}_{pppp} \\ &= \left(\delta_{ai}\delta_{jb}\delta_{kb}\delta_{\ell b} + \delta_{ib}\delta_{aj}\delta_{kb}\delta_{\ell b} + \delta_{ib}\delta_{jb}\delta_{ak}\delta_{\ell b} + \delta_{ib}\delta_{jb}\delta_{kb}\delta_{a\ell} \right) \mathscr{S}_{bbbb}.\end{aligned}$$

Subsequently, we have that

$$\left\| \Gamma^{ab} \operatorname{vec}\left(\mathscr{S}\right) \Delta\mathbf{Q}_{ab} \right\|_F^2 = \sum_{i,j,k,\ell=1}^{n} \left| \left(\delta_{ai}\delta_{jb}\delta_{kb}\delta_{\ell b} + \delta_{ib}\delta_{aj}\delta_{kb}\delta_{\ell b} + \delta_{ib}\delta_{jb}\delta_{ak}\delta_{\ell b} + \delta_{ib}\delta_{jb}\delta_{kb}\delta_{a\ell} \right) \mathscr{S}_{bbbb}\Delta\mathbf{Q}_{ab} \right|^2.$$

It can be easily shown that for $a = b$

$$\left\| \Gamma^{bb} \operatorname{vec}\left(\mathscr{S}\right) \Delta\mathbf{Q}_{bb} \right\|_F^2 = 16 \left| \mathscr{S}_{bbbb}\Delta\mathbf{Q}_{bb} \right|^2,$$

and for $a \neq b$

$$\left\| \Gamma^{ab} \operatorname{vec}\left(\mathscr{S}\right) \Delta\mathbf{Q}_{ab} \right\|_F^2 = 4 \left| \mathscr{S}_{bbbb}\Delta\mathbf{Q}_{ab} \right|^2.$$

We therefore have

$$\begin{aligned}(\star) &= \sum_{i=j}\left\|\frac{\partial\tilde{f}}{\partial\mathbf{Q}_{ii}}(u)^T\Delta\mathbf{Q}_{ii}\right\|_F^2+\sum_{i\neq j}\left\|\frac{\partial\tilde{f}}{\partial\mathbf{Q}_{ij}}(u)^T\Delta\mathbf{Q}_{ij}\right\|_F^2+\|\Delta\mathscr{S}\|_2^2\\ &= 16\sum_{i=j}|\mathscr{S}_{iiii}|^2\,|\Delta\mathbf{Q}_{ii}|^2+4\sum_{i\neq j}|\mathscr{S}_{iiii}|^2\,|\Delta\mathbf{Q}_{ij}|^2+\|\Delta\mathscr{S}\|_2^2\\ &\geq 4\sum_{i,j}|\mathscr{S}_{iiii}|^2\,|\Delta\mathbf{Q}_{ij}|^2+\|\Delta\mathscr{S}\|_2^2\\ &= (\star)\end{aligned}$$

Now assume that $|\mathscr{S}_{iiii}|\geq\epsilon_{\mathscr{S}}$ for all i, therefore

$$\begin{aligned}(\star)&\geq 4\epsilon_{\mathscr{S}}^2\|\Delta\mathbf{Q}\|_F^2+\|\Delta\mathscr{S}\|_2^2\\ &\geq 4\epsilon_{\mathscr{S}}^2\|\Delta u\|_2^2.\end{aligned}$$

In the last line, we assume w.l.o.g that $4\epsilon_{\mathscr{S}}^2\leq 1$. We have therefore proved that

$$\min_{\|\Delta u\|=1}\left\|D\tilde{f}\,(u)^T\,\Delta u\right\|_2\geq 2\epsilon_{\mathscr{S}}, \tag{3.68}$$

yielding $C_3\coloneqq 2\epsilon_{\mathscr{S}}$. □

We have the following lemma to cover the set of short secants.

Lemma 8 (Short Secants Covering Number)**.** *Let $\underline{\Omega'}=\{u_i\}$ be an ϵ- cover for $\Omega'=O(n)\times(\mathfrak{D}\cap\mathfrak{B}_1(0))$ and considering the following:*

1. $\left\|f(u)-f(u')-Df_{u'}^T(u-u')\right\|\leq C_1\|u-u'\|^2$ *(Taylor approximation Lemma 5)*
2. $\|Df_u-Df_{u'}\|\leq C_2\|u-u'\|$ *(bounded curvature Lemma 6)*
3. $\left\|Df_u^\dagger\right\|\leq C_3$ *(bounded gradient Lemma 7),*

where $f:\Omega\mapsto\mathfrak{S}_\theta$ is defined in Eqn. 3.45 and Df_u defines the first order differential of f evaluated at the point u. Then given $u_i\in\Omega$, $\forall u,u'\in\mathfrak{B}_{\epsilon_0}(u_i)$ and $\|\mathscr{Z}-\mathscr{Z}'\|\leq\eta$, where $\mathscr{Z}=f(u)$ and $\mathscr{Z}'=f(u')$, we have

$$\left\|\frac{\mathscr{Z}-\mathscr{Z}'}{\|\mathscr{Z}-\mathscr{Z}'\|}-Df_{u_i}^T\frac{u-u'}{\|\mathscr{Z}-\mathscr{Z}'\|}\right\|\leq C_4\epsilon_0. \tag{3.69}$$

where $C_4 \coloneqq C_3(2C_1 + C_2)$.

Proof.

$$\begin{aligned}\left\|\mathscr{Z} - \mathscr{Z}' - Df_{u_i}^T(u-u')\right\| &= \left\|f(u) - f(u') - Df_u^T(u-u') + \left(Df_u - Df_{u_i}\right)^T\left(u-u'\right)\right\| \\ &\leq \left\|f(u) - f(u') - Df_u^T(u-u')\right\| + \left\|\left(Df_u - Df_{u_i}\right)^T\left(u-u'\right)\right\| \\ &\leq C_1\left\|u-u'\right\|^2 + C_2\left\|u-u_i\right\|\left\|u-u'\right\| \\ &= (\star)\end{aligned}$$

Given that $u, u' \in \mathfrak{B}_{\epsilon_0}(u_i)$, we have that $\|u-u_i\| \leq \epsilon_0$ and $\|u-u'\| \leq 2\epsilon_0$. Therefore

$$\begin{aligned}(\star) &\leq 2C_1\epsilon_0\left\|u-u'\right\| + C_2\epsilon_0\left\|u-u'\right\| \\ &= (2C_1 + C_2)\epsilon_0\left\|u-u'\right\|.\end{aligned}$$

Now dividing by $\|\mathscr{Z} - \mathscr{Z}'\|$ gives:

$$\begin{aligned}\left\|\frac{\mathscr{Z}-\mathscr{Z}'}{\|\mathscr{Z}-\mathscr{Z}'\|} - Df_{u_i}^T\frac{u-u'}{\|\mathscr{Z}-\mathscr{Z}'\|}\right\| &\leq (2C_1 + C_2)\frac{\|u-u'\|}{\|\mathscr{Z}-\mathscr{Z}'\|} \\ &\leq C_3(2C_1 + C_2).\end{aligned}$$

In the last line, we have used the fact that bounded (inverse) gradient implies Lipschitzness. □

As a result, the set of bounded tangent vectors, defined by

$$\mathcal{V} \coloneqq \left\{Df_{u_i}^T\frac{u-u'}{\|\mathscr{Z}-\mathscr{Z}'\|} \mid \forall u_i \in \Omega\right\} \tag{3.70}$$

forms a generalized ϵ-cover for $\mathfrak{N}_\eta^{\mathfrak{c}}\left(\bar{\mathfrak{S}}_\theta - \bar{\bar{\mathfrak{S}}}_\theta\right)$ with covering number (see Lemma 4.3 of [97])

$$\begin{aligned}\mathrm{CN}\left(\mathcal{V}, \|\cdot\|_F, \epsilon\right) &\leq C_4\,\mathrm{CN}\left(\bar{\bar{\mathfrak{S}}}_\theta, \|\cdot\|_F, \epsilon_0\right)\left(\frac{3}{\epsilon}\right)^{\frac{n(n+1)}{2}} \\ &\leq C_4\left(\frac{6}{\epsilon_0}\right)^{n(n+1)}\left(\frac{3}{\epsilon}\right)^{\frac{n(n+1)}{2}}.\end{aligned}$$

From Eqn 3.48, we can bound the covering number of the normalized secant set:

$$\begin{aligned}
\text{CN}\left(\mathfrak{N}\left(\mathfrak{S}_\theta - \mathfrak{S}_\theta\right), \|\cdot\|_F, \epsilon\right) &\leq \text{CN}\left(\mathfrak{N}_\eta\left(\bar{\mathfrak{S}}_\theta - \bar{\mathfrak{S}}_\theta\right), \|\cdot\|_F, \epsilon\right) \;+\; \text{CN}\left(\mathfrak{N}_\eta^c\left(\bar{\mathfrak{S}}_\theta - \bar{\bar{\mathfrak{S}}}_\theta\right), \|\cdot\|_F, \epsilon\right) \\
&\leq \left(\frac{6}{\gamma\epsilon}\right)^{2n(n+1)} \;+\; C_4\left(\frac{6}{\epsilon_0}\right)^{n(n+1)}\left(\frac{3}{\epsilon}\right)^{\frac{n(n+1)}{2}} \\
&\leq \left(\frac{6}{\gamma\epsilon}\right)^{2n(n+1)} \;+\; C_4\left(\frac{6}{\epsilon_0}\right)^{n(n+1)}\left(\frac{3}{\epsilon}\right)^{n(n+1)} \\
&= \left(\frac{6}{\gamma\epsilon}\right)^{2n(n+1)} \;+\; C_4\left(\frac{18}{\epsilon_0\epsilon}\right)^{n(n+1)} \\
&= (\star).
\end{aligned}$$

Note that by definition $\epsilon_0 \leq \eta\ (= 4\gamma)$, therefore $\gamma \geq \frac{\epsilon_0}{4}$. As a result

$$\begin{aligned}
(\star) &\leq C_4\left(\left(\frac{24}{\epsilon_0\epsilon}\right)^{2n(n+1)} \;+\; \left(\frac{24}{\epsilon_0\epsilon}\right)^{n(n+1)}\right) \\
&\leq \left(\frac{48C_4}{\epsilon_0\epsilon}\right)^{2n(n+1)} \\
&\leq \left(\frac{C_0}{\epsilon}\right)^{2n(n+1)}
\end{aligned}$$

where $C_0 = \frac{48C_4}{\epsilon_0}$.

3.A.2 Proof of Theorem 2

Proof. First, note that as the Frobenius norm is rotationally invariant we have that

$$\|\mathscr{Z}\|_F = \|\mathscr{S} \times_1 \mathbf{Q} \times_2 \cdots \times_4 \mathbf{Q}\|_F = \|\mathscr{S}\|_F \leq R.$$

As $\hat{\mathscr{Z}}$ is an empirical average of the true expected cumulant tensor $\mathscr{Z}$, we can use a version of the vectorial Hoeffding's inequality in Lemma 4 of [100] that states with probability at least $1-\rho$ on the random draw of $\mathbf{z}_1, \ldots, \mathbf{z}_N$ that

$$\|\hat{\mathscr{Z}} - \mathscr{Z}\|_F \leq \frac{R\left(1 + \sqrt{2\log(1/\rho)}\right)}{\sqrt{N}}. \tag{3.71}$$

Next, we can use the boundedness property of random Gaussian measurements [101]. First, let

$\mathbf{e} = \text{vec}\left(\hat{\mathscr{Z}} - \mathscr{Z}\right)$ denote the finite approximation error from above. Then the boundedness property of subgaussian matrices (see Definition 6.2 in [101]) states with probability at least $1 - \xi$ on the sampling of $\mathbf{A}$, that

$$\|\mathbf{A}\mathbf{e}\|_2 \leq C\|\mathbf{e}\|_2 \tag{3.72}$$

for some constant $C > 0$.

Combining the two equations, we get with probability at least $(1-\rho)(1-\xi) \geq 1 - \rho - \xi$ on the drawing of both $\mathbf{A}$ and $\mathbf{z}_1, \ldots, \mathbf{z}_N$ that

$$\|\mathcal{A}(\mathscr{Z}) - \mathcal{A}(\hat{\mathscr{Z}})\|_2 \leq \frac{CR\left(1 + \sqrt{2\log(1/\rho)}\right)}{\sqrt{N}}. \tag{3.73}$$

□

3.A.3 Conic Properties

Definition 9 (Conic Set)**.** *A set S is conic if for all $x \in S$ the positive scalar multiple $\alpha x \in S$ for $\alpha > 0$.*

Lemma 9. *The ICA model set $\mathfrak{S}_\theta$ is conic*

Proof. Let $\mathscr{Z} \in \mathfrak{S}_\theta$ have decomposition $\mathscr{Z} = \mathscr{S} \times_1 \mathbf{Q} \times_2 \mathbf{Q} \times_3 \mathbf{Q} \times_4 \mathbf{Q}$. For any scalar $\alpha > 0$ we show that $\alpha\mathscr{Z} \in \mathfrak{S}_\theta$:

$$\begin{aligned}\alpha\mathscr{Z} =& \alpha\left(\mathscr{S} \times_1 \mathbf{Q} \times_2 \mathbf{Q} \times_3 \mathbf{Q} \times_4 \mathbf{Q}\right) \\ =& \left(\alpha\mathscr{S}\right) \times_1 \mathbf{Q} \times_2 \mathbf{Q} \times_3 \mathbf{Q} \times_4 \mathbf{Q}.\end{aligned}$$

As positive scalar multiplication does not change the support of $\mathscr{S}$ then $\alpha\mathscr{S} \in \mathfrak{D}$. Let $\hat{\mathscr{S}} := \alpha\mathscr{S}$, then

$$\alpha\mathscr{Z} = \hat{\mathscr{S}} \times_1 \mathbf{Q} \times_2 \mathbf{Q} \times_3 \mathbf{Q} \times_4 \mathbf{Q}, \tag{3.74}$$

and therefore $\alpha\mathscr{Z} \in \mathfrak{S}_\theta$ for all $\alpha > 0$. □

3.A.4 1st Order Differential

Define $\tilde{f}(u) = \mathbf{Q} \otimes \mathbf{Q} \otimes \mathbf{Q} \otimes \mathbf{Q} \operatorname{vec}(\mathscr{S})$, where $u = (\mathbf{Q}, \mathscr{S})$. We can decompose the 1st order differential as

$$D\tilde{f}(u) = \left[\frac{\partial \tilde{f}}{\partial \mathbf{Q}}(u), \frac{\partial \tilde{f}}{\partial \mathscr{S}}(u) \right]. \tag{3.75}$$

We begin with the left hand side of the equation above. Recall $\bar{\mathbf{Q}} = \mathbf{Q} \otimes \mathbf{Q} \otimes \mathbf{Q} \otimes \mathbf{Q}$, then

$$\begin{aligned} \Gamma_{ij} &= \frac{\partial \bar{\mathbf{Q}}}{\partial \mathbf{Q}_{ij}} \\ &= \mathbf{E}^{ij} \otimes \mathbf{Q} \otimes \mathbf{Q} \otimes \mathbf{Q} + \mathbf{Q} \otimes \mathbf{E}^{ij} \otimes \mathbf{Q} \otimes \mathbf{Q} \\ &\quad + \mathbf{Q} \otimes \mathbf{Q} \otimes \mathbf{E}^{ij} \otimes \mathbf{Q} + \mathbf{Q} \otimes \otimes \mathbf{Q} \otimes \mathbf{Q} \otimes \mathbf{E}^{ij}. \end{aligned}$$

We therefore have that

$$\frac{\partial \tilde{f}}{\partial \mathbf{Q}_{ij}}(u) = \Gamma_{ij} \operatorname{vec}(\mathscr{S}). \tag{3.76}$$

Recalling that $\mathscr{S}_{ijkl} = 0$ for all $ijkl \neq iiii$, we therefore have that

$$\frac{\partial \tilde{f}}{\partial \mathscr{S}_{kkkk}}(u) = \bar{\mathbf{Q}} e_{\hat{k}},$$

where $\hat{k}$ is the equivalent vectorised position of index $kkkk$.

3.A.5 Second Order Differential

We begin by decomposing the 2nd order differential into

$$H\tilde{f}(u) = \begin{bmatrix} \dfrac{\partial^2 \tilde{f}}{\partial \mathbf{Q} \partial \mathbf{Q}}(u), & \dfrac{\partial^2 \tilde{f}}{\partial \mathbf{Q} \partial \mathscr{S}}(u) \\ \dfrac{\partial^2 \tilde{f}}{\partial \mathbf{Q} \partial \mathscr{S}}(u), & \dfrac{\partial^2 \tilde{f}}{\partial \mathscr{S} \partial \mathscr{S}}(u) \end{bmatrix}. \tag{3.77}$$

Firstly, given

$$\begin{aligned}\Pi_{ijk\ell} \;&=\; \frac{\partial^2 \bar{\mathbf{Q}}}{\partial \mathbf{Q}_{ij} \partial \mathbf{Q}_{k\ell}} \\ &=\; \mathbf{E}^{ij} \otimes \mathbb{E}^{k\ell} \otimes \mathbf{Q} \otimes \mathbf{Q} \;+\; \mathbf{E}^{ij} \otimes \mathbf{Q} \otimes \mathbb{E}^{k\ell} \otimes \mathbf{Q} \;+\; \mathbf{E}^{ij} \otimes \mathbf{Q} \otimes \mathbf{Q} \otimes \mathbb{E}^{k\ell} \\ &\quad+\; \mathbb{E}^{k\ell} \otimes \mathbf{E}^{ij} \otimes \mathbf{Q} \otimes \mathbf{Q} \;+\; \mathbf{Q} \otimes \mathbf{E}^{ij} \otimes \mathbb{E}^{k\ell} \otimes \mathbf{Q} \;+\; \mathbf{Q} \otimes \mathbf{E}^{ij} \otimes \mathbf{Q} \otimes \mathbb{E}^{k\ell} \\ &\quad+\; \mathbb{E}^{k\ell} \otimes \mathbf{Q} \otimes \mathbf{E}^{ij} \otimes \mathbf{Q} \;+\; \mathbf{Q} \otimes \mathbb{E}^{k\ell} \otimes \mathbf{E}^{ij} \otimes \mathbf{Q} \;+\; \mathbf{Q} \otimes \mathbf{Q} \otimes \mathbf{E}^{ij} \otimes \mathbb{E}^{k\ell} \\ &\quad+\; \mathbb{E}^{k\ell} \otimes \mathbf{Q} \otimes \mathbf{Q} \otimes \mathbf{E}^{ij} \;+\; \mathbf{Q} \otimes \mathbb{E}^{k\ell} \otimes \mathbf{Q} \otimes \mathbf{E}^{ij} \;+\; \mathbf{Q} \otimes \mathbf{Q} \otimes \mathbb{E}^{k\ell} \otimes \mathbf{E}^{ij}\end{aligned}$$

we therefore have that

$$\frac{\partial^2 \tilde{f}}{\partial \mathbf{Q} \partial \mathbf{Q}}(u) \;=\; \Pi ijk\ell \,\mathrm{vec}\,(\mathscr{S})\,. \tag{3.78}$$

Secondly, it should be straightforward to see that

$$\frac{\partial^2 \tilde{f}}{\partial \mathbf{Q} \partial \mathscr{S}}(u) = \Gamma_{ij} e_{\hat{k}}. \tag{3.79}$$

Finally, we have that

$$\frac{\partial^2 \tilde{f}}{\partial \mathscr{S}_{iiii} \partial \mathscr{S}_{jjjj}} \;=\; \mathbf{0}_{n^4 \times 1}, \tag{3.80}$$

where $\mathbf{0}_{n^4 \times 1}$ is zero vector of length $n^4 \times 1$.

Chapter 4
Compressive Learning for Semi-Parametric Models

4.1 Introduction

In the previous chapter, we built a compressive learning framework for ICA that enables substantial compression compared to existing ICA algorithms. As the ICA is typically left distribution free, it was showed that one could exploit particular structure of the data's cumulants to firstly construct a dimensionality reducing sketch and, secondly, design tractable recovery algorithms. ICA is part of the wider class of semi-parametric models that were introduced in Section 2.1.4 that share the inherent characteristic of only being partially parametrized. The overriding consequence is that the conventional sketch, that maps from the space of distributions, $\mathcal{A} : \mathcal{P}(\mathcal{X}) \mapsto \mathbb{C}^m$ defined in Eqn 2.61, is difficult to design. As is the case for both the compressive PCA and ICA models introduced in Section 2.3.3 and 3.2, respectively, one must leverage some intermediary statistic that permits identifiability of the model to build the associated sketch.

The compressive learning framework originally developed by Gribonval et al. in [1, 6] and discussed in Section 2.3 was designed primarily for parametric models and assumes the existence of a sketching operator $\mathcal{A}$ that maps from the space of distributions $\mathcal{P}(\mathcal{X})$. As such, it does not provide a straightforward blueprint to help design both practical sketches and tractable algorithms. In this chapter, we reformulate the existing compressive learning framework to explicitly cater for semi-parametric models. The resulting reformulation enables practitioners to identify and potentially exploit the properties of semi-parametric models to help build efficient compressive learning schemes.

Below we highlight the main contributions of this chapter.

- By taking into account the unique characteristics of semi-parametric models, we reformulate the existing compressive learning framework to cater explicitly for such distribution-free models. In the reformulation, the sketch $\mathcal{A}$ maps not from the space of distribution

but some intermediate statistic space. As a result, the decoder recovers an object from the intermediate statistic space that achieves minimum expect risk with respect to the model.

- A compressive generalized PCA scheme is introduced and, utilizing the reformulated framework, we examine and compare it to the compressive ICA scheme developed in Chapter 3 demonstrating when compression is and is not attainable. In this particular case study, the reformulated compressive learning framework identifies when efficient compressive learning is possible with respect to the dimensions of the model which might not have been as easily detected by the existing formulation.

- We highlight the unique theoretical and practical advantages and disadvantages of building a compressive learning scheme for a semi-parametric model in comparison to a fully parametrized model. For instance, as we leverage some intermediate finite statistic to construct our sketch, the compressive learning problem typically reduces to a generalized finite dimensional compressive sensing problem. One can therefore leverage the long-establish theory and techniques of compressive sensing to help develop theoretical gaurantees and tractable decoders.

The rest of this chapter is organized as follows: In Section 4.2 we highlight the unique characteristics of semi-parametric models that are not explicitly catered for in the existing compressive learning framework. Section 4.3 reformulates the original compressive learning framework to be explicit in the construction of a compressive semi-parametric scheme. In Section 4.4, we provide a case study on when compressive semi-parametric learning succeeds and fails by introducing a compressive generalized PCA scheme and comparing attainable compression with the CICA scheme introduced in Chapter 3. A discussion of the advantages and disadvantages of compressive semi-parametric learning are detailed in Section 4.5 and then conclude the chapter in Section 4.6 with some final remarks.

4.2 Motivation

Figure 4.1 depicts a schematic of the original compressive learning framework proposed in [1].

For convenience, we recall the main properties of the framework, however see Section 2.3 for

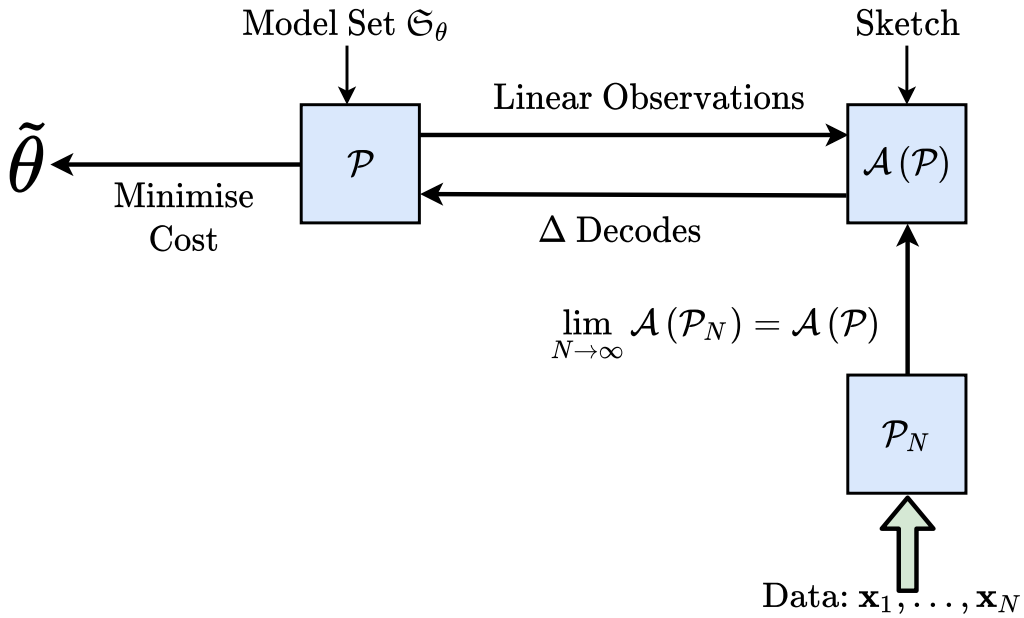

Figure 4.1: *A schematic diagram detailing the original compressive learning framework proposed in [1].*

a detailed discussion. One typically designs a sketching operator $\mathcal{A} : \mathcal{P}(\mathcal{X}) \mapsto \mathbb{C}^m$ such that

$$\mathbf{y} \;=\; \mathcal{A}(\mathcal{P}) \;=\; \mathbb{E}_{\mathbf{x}\sim\mathcal{P}}\Phi(\mathbf{x})\,. \tag{4.1}$$

However, owing to finite data sample we form an empirical sketch

$$\mathbf{y}_N \;=\; \mathcal{A}(\mathcal{P}_N) \;=\; \frac{1}{N}\sum_{i=1}^{N}\Phi(\mathbf{x}_i)\,. \tag{4.2}$$

Hence, given an empirical sketch $\mathbf{y}_N$, the goal is to recover the parameters of the distribution via decoder, for example

$$\tilde{\theta} \;\in\; \Delta(\mathbf{y}_N, \mathcal{A})\,. \tag{4.3}$$

Typically, the decoder is made tractable by assuming some regularity assumptions to the set of solutions of the learning task. Such assumptions are formalised via the model set of the learning problem which is defined by

$$\mathfrak{S}_\theta \;=\; \{\mathcal{P} \in \mathcal{P}(\mathcal{X}) \mid \exists\theta \in \Theta, \mathcal{R}(\theta, \mathcal{P}) = 0\}\,. \tag{4.4}$$

In other words, we restrict the set of solutions to distributions that achieve zero expected risk.

As shown in Section 2.3, the cost function associated with the compressive learning model typically reduces to

$$\mathcal{P}_{\tilde{\theta}} \;=\; \underset{\mathcal{P}\in\mathfrak{S}_\theta}{\arg\min}\|\mathbf{y}_N - \mathcal{A}\left(\mathcal{P}\right)\|_2^2. \tag{4.5}$$

The decoder in Eqn 4.5 assumes the existence of some distributional form of the model. However, in the case of semi-parametric learning tasks the model is not fully parameterized.

4.2.1 Topology of Semi-Parametric Models

Semi-parametric models form an interesting class of models which are used extensively in machine and statistical learning tasks. In its crudest form, a semi-parametric model is one that has no parametric constraints on the data distribution but is identifiable via a particular set of statistics of the data distribution. A semi-parametric model can be formally described by θ, together with a function $g \in \mathcal{G}$, such that the model is specified by the set $\mathcal{P}\left(\mathcal{X}\right), \Theta, \mathcal{G}$ and the parametrization given by [11]:

$$(\theta, g) \;\mapsto \mathcal{P}_{(\theta,g)} \quad \text{for} \;\; (\theta, g) \in \Theta \times \mathcal{G}. \tag{4.6}$$

In most cases, the function g is not known a-priori nor is it sufficiently smooth to approximate [11]. Subsequently, gaining explicit access to the space of distributions is intractable. In many instances, one can use some particular set of statistics to both identify and solve the semi-parametric task. This is demonstrated in PCA and ICA where the covariance matrix and cumulant tensor of the data can be used to identify each model, respectively. Throughout this chapter we will term such statistics, which are used to solve the semi-parametric problem, as *identifiable statistics.*

Let $\Sigma_{\mathcal{P}}$ denote an identifiable statistic associated with an arbitrary semi-parametric model. An equivalence exists between distributions in the model set and the set of identifiable statistics which we denote by $\mathcal{S}\left(\mathcal{X}\right)$. Formally, let $\sim$ be the equivalence relation defined by

$$\mathcal{P} \sim \mathcal{Q} \quad \text{iff} \quad \Sigma_{\mathcal{P}} = \Sigma_{\mathcal{Q}} \tag{4.7}$$

for $\mathcal{P}, \mathcal{Q} \in \mathfrak{S}_\theta$. Due to the equivalence relation, there exists a many-to-one mapping defined

by

$$\vartheta : \mathcal{P}(\mathcal{X}) \mapsto \mathcal{S}(\mathcal{X}) \tag{4.8}$$

that maps the equivalence classes in the space of distributions $\mathcal{P}(\mathcal{X})$ to a single point in the set $\mathcal{S}(\mathcal{X})$. Figure 4.2, illustrates the equivalence class structure of semi-parametric models and the mapping ϑ to the set of identifiable statistics $\mathcal{S}(\mathcal{X})$.

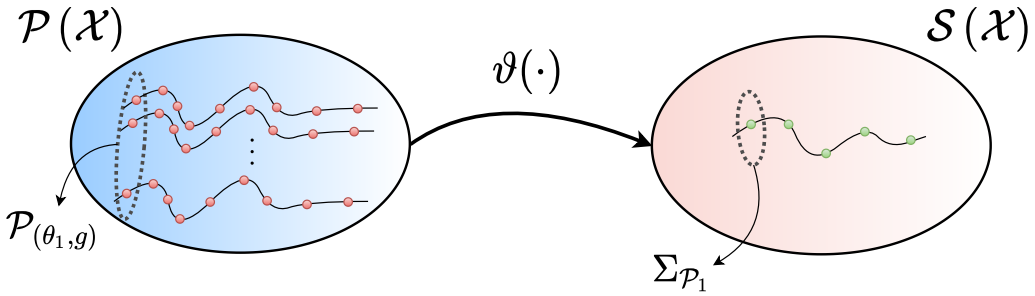

Figure 4.2: *A schematic diagram of the probability equivalence class whereby many distributions collapse down to one point on the intermediary set of statistics.*

Due to the equivalence class structure that is inherent in semi-parametric models, we lose the luxury of injectivity between $\theta \mapsto \mathcal{P}$ that exists for parametric models. As a result, a single distribution cannot straightforwardly be encoded by a sketch and then subsequently recovered via a decoder.

4.3 Compressive Learning Framework for Semi-Parametric Models

To reformulate the original compressive learning framework, we assume the existence of an identifiable statistic $\Sigma_{\mathcal{P}} \in \mathcal{S}(\mathcal{X})$ calculated over the data $\mathcal{X}$ that can be utilized to identify and estimate the parameters of the semi-parametric model and can therefore be used to define a risk function. To start, we define the risk function over each equivalence class. This is possible when there exists a map $\vartheta : \mathcal{P}(\mathcal{X}) \mapsto \mathcal{S}(\mathcal{X})$ satisfying

$$\mathcal{R}(\mathcal{P}, \theta) \;=\; \mathcal{R}(\vartheta(\mathcal{P}), \theta). \tag{4.9}$$

As a consequence of Eqn 4.9, the parametrization of the probability distribution is not always required as it often suffices to have a parametrization of the statistic set $\mathcal{S}(\mathcal{X})$. In accordance,

we can define the semi-parametric sketch as

$$\mathcal{A}\left(\Sigma_{\mathcal{P}}\right) \;=\; \mathcal{A}\left(\vartheta\left(\mathcal{P}\right)\right) \;=\; \mathbb{E}_{\mathbf{x}\sim\vartheta(\mathcal{P})}\Phi\left(\mathbf{x}\right) \tag{4.10}$$

where $\mathcal{A} : \mathcal{S}\left(\mathcal{X}\right) \mapsto \mathbb{C}^m$ defines the semi-parametric sketching operator.

The difference between the sketching operator defined in Eqn 4.10 and the original sketching operator in Eqn 2.61 may look at first sight quite subtle. However, notice that the sketching operator $\mathcal{A}$ is now an operator acting over the space of identifiable statistics $\mathcal{S}\left(\mathcal{X}\right)$ instead of the space of probability distributions $\mathcal{P}\left(\mathcal{X}\right)$. Moreover, the space of identifiable statistics is typically finite dimensional which is in contrast to the infinite dimensional probability space.

Remark 5. *Eqn 4.10 shows that the sketch is equal to the expectation of a feature function with respect to the equivalence class of distributions. However, in practice we typically don't have access to either $\mathcal{P}$ or $\vartheta\left(\mathcal{P}\right)$ and therefore computing $\mathbb{E}_{\mathbf{x}\sim\vartheta(\mathcal{P})}\Phi\left(\mathbf{x}\right)$ is difficult and often intractable. As will be shown in Section 4.4, we exploit structural redundancies of the set of identifiable statistics $\mathcal{S}\left(\mathcal{X}\right)$ to construct a tractable inference scheme.*

Owing to finite data samples, we define the empirical sketch $\mathbf{y}_N$ as the sketch computed over the finite sample statistic approximation Σ_N, for example

$$\mathbf{y}_N \;=\; \mathcal{A}\left(\Sigma_N\right) \;=\; \frac{1}{N}\sum_{i=1}^{N}\Phi\left(\mathbf{x}_i\right). \tag{4.11}$$

Due to the law of large numbers, the finite approximated statistic approaches the true statistic, i.e. $\lim_{N\to\infty}\Sigma_N = \Sigma_{\mathcal{P}}$.

Provided a sketch has been computed, one must recover the sketched object to enable the estimation of the models parameters θ. However, in the case of compressive semi-parametric learning, the object of interest is a finite dimensional statistic $\Sigma_{\mathcal{P}}$ instead of a parametrized distribution $\mathcal{P}$. In general, the regularity assumptions of the model set $\mathfrak{S}_\theta$ manifest into structural redundancies of the statistic $\Sigma_{\mathcal{P}}$. For example, recall the compressive PCA scheme detailed in Section 2.3.3, whereby the model set is defined as

$$\mathfrak{S}_k \;=\; \left\{\mathcal{P} \mid \text{rank}\left(\mathbf{C}_{\mathbf{x}}\right) \leq K\right\}. \tag{4.12}$$

In other words, the set of distributions that are supported on a K-dimensional orthonormal basis

induces the rank of the second order moment statistic $\Sigma_{\mathcal{P}} = \mathbf{C}_{\mathbf{x}}$ to be at most rank K. Similarly, we demonstrated in Chapter 3 that solutions to the ICA learning task admit a cumulant tensor $\Sigma_{\mathcal{P}} = \mathscr{K}$ that exhibits a sparse tensor decomposition. In such circumstances, one can design decoders Δ that recover a estimate $\tilde{\Sigma}_{\mathcal{P}}$ by exploiting structural redundancies, i.e.

$$\tilde{\Sigma}_{\mathcal{P}} \in \Delta\left(\mathbf{y}_N, \mathcal{A}\right). \tag{4.13}$$

Figure 4.3 details the newly formulated compressive learning framework for semi-parametric models.

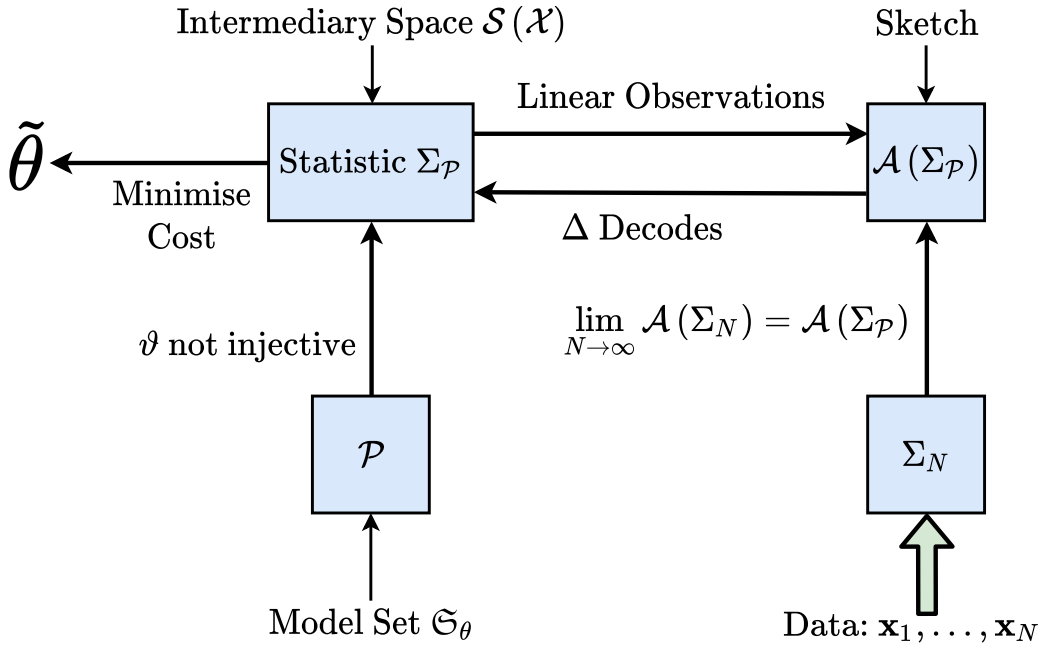

Figure 4.3: *A schematic diagram detailing the reformulated compressive learning framework for semi-parametric models.*

4.4 When Does Compressive Semi-Parametric Learning Work? - A Case Study

In this Section, we provide a case study between two compressive semi-parametric schemes that either succeed or fail to lead to substantial compression with respect to the full data $\mathbf{X}$. First, we introduce the compressive generalized principal component analysis model.

4.4.1 Compressive Generalized Principal Component Analysis

We analyse a compressive subspace clustering scheme[1] through the lens of the reformulated compressive learning framework proposed in Section 4.3. First, we briefly detail the fundamental concepts of subspace clustering and refer the interested reader to [103, 104] for a thorough exposition. The subspace clustering problem consists of finding the best union of subspaces that fits a given dataset. Let us denote by $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N \in \mathbb{R}^d$ the finite dataset, then subspace clustering attempts to find n subspaces $S_i \subset \mathbb{R}^d$ for $i = 1, \ldots, n$ where the dimension of each subspace is denoted by $\dim(S_i) = d_i$. To make the learning task well-posed, we assume that the number of subspaces and the associated dimensions are known in advance. The task of subspace clustering can be seen of as an n-mixture model with data sampled from n unknown probability distributions $\mathcal{P}_i$. It is therefore straightforward to see that subspace clustering falls within the class of semi-parametric models described in Section 4.3. Figure 4.4 depicts an example of a set of data points that lie on the union of a plane ($d_1 = 2$) and 2 lines ($d_2 = d_3 = 1$) sampled from $n = 3$ unknown probability distributions. Subsequently, the model set to the subspace clustering problem can be defined as

$$\mathfrak{S}_\theta \coloneqq \left\{ \mathcal{P} = \sum_i \alpha_i \mathcal{P}_{h_i} \mid \text{rank}(\Sigma_{\mathcal{P}_{h_i}}) \leq d_i, \sum_i \alpha_i = 1 \text{ and } \alpha_i > 0 \right\}. \tag{4.14}$$

The subspace clustering problem has previously been solved through a generalised principle component analysis (GPCA) approach [103]. Consider the specific case where the data are distributed according to a union of two planes in $\mathbb{R}^3$, each one with normal vector $\mathbf{b}_i \in \mathbb{R}^3$. The union of two planes can be expressed as a set of points [104] such that

$$p(\mathbf{x}) \;=\; (\mathbf{b}_1^T \mathbf{x})(\mathbf{b}_2^T \mathbf{x}) \;=\; 0. \tag{4.15}$$

This equation can be reduced to the equation of a conic of the form

$$c_1 x_1^2 + c_2 x_1 x_2 + c_3 x_1 x_3 + c_4 x_2^2 + c_5 x_2 x_3 + c_6 x_3^2 \;=\; 0. \tag{4.16}$$

More generally, data drawn from the union of n $d-1$ subspaces of $\mathbb{R}^d$ can be represented by

[1] This section is based on the paper [102]. It should be noted that the compressive GPCA scheme was proposed by Antoine Gonon in an internship circa 2018.

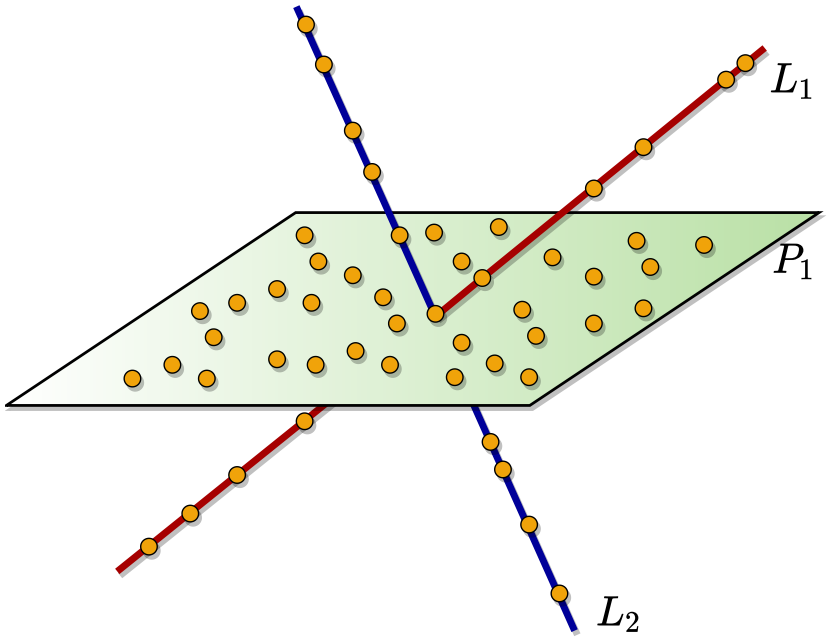

Figure 4.4: *A set of data points in $\mathbb{R}^3$ drawn from a mixture of three distributions that supported on the plane P_1 ($d_1 = 2$), the line L_1 ($d_2 = 1$) and the line L_2 ($d_3 = 1$).*

the polynomial of the form

$$p(\mathbf{x}) \;=\; (\mathbf{b}_1^T\mathbf{x})(\mathbf{b}_2^T\mathbf{x})\dots(\mathbf{b}_n^T\mathbf{x}) \;=\; 0, \tag{4.17}$$

where the vector $\mathbf{b}_i \in \mathbb{R}^d$ is orthogonal to the corresponding subspace S_i. This polynomial is of degree n in $\mathbf{x}$ and can be written as $\mathbf{c}^T\nu_n(\mathbf{x})$ where $\mathbf{c}$ is the vector of coefficients and $\nu_n(\mathbf{x})$ is the Veronese embedding containing all the distinct monomials of degree n in $\mathbf{x}$. The embedded point $\nu_n(\mathbf{x})$ belong to $\mathbb{R}^D$ where

$$D \;\coloneqq\; \binom{n+d-1}{d-1} \leq d^n. \tag{4.18}$$

In the case of noiseless data, the vector of coefficients $\mathbf{c}$ of each polynomial can be computed from

$$[p(\mathbf{x}_1),\dots,p(\mathbf{x}_N)] = \mathbf{c}^T\,[\nu_n(\mathbf{x}_1),\dots,\nu_n(\mathbf{x}_N)] \;\coloneqq\; \mathbf{c}^T\mathbf{V} \;=\; \mathbf{0} \tag{4.19}$$

and the number of polynomials is simply the dimension of the null space of $\mathbf{V}^T$. The relationship between the number of subspaces n, their dimension d_i, the number of polynomi-

als and therefore the null space dimension of $\mathbf{V}$ involves the theory of Hilbert functions (see [105, 106]). Without going into details, we assume in this chapter that the specific makeup of the subspaces, for instance n and d_i, directly characterises the dimension of the null space of $\mathbf{V}$. Indeed, computing the null space of $\mathbf{V}$ can be easily deduced by finding the eigendecomposition of the correlation matrix $\mathbf{C}_\nu \in \mathbb{R}^{D\times D}$ of the embedded data:

$$\mathbf{C}_\nu \coloneqq \mathbb{E}_{\mathbf{x}\sim\mathcal{P}}\nu_n(\mathbf{x})\nu_n(\mathbf{x})^T \approx \frac{1}{N}\mathbf{V}\mathbf{V}^T. \tag{4.20}$$

Once the coefficients $\mathbf{c}$ have been determined, we can retrieve each normal vector $\mathbf{b}_i$ directly from $\mathbf{c}$. This can be done by taking the derivatives of the polynomials at a data point $\mathbf{x}$. In the case of n subspaces, we have $p(\mathbf{x}) = (\mathbf{b}_1^T\mathbf{x})\dots(\mathbf{b}_n^T\mathbf{x})$ and $\nabla p(\mathbf{x}) \sim \mathbf{b}_i$ if $\mathbf{x} \in S_i$. Using this result, we can obtain a set of all normal vectors to $\mathbf{S}_i$ from the derivatives of all the polynomials at $\mathbf{x} \in S_i$. This allows us to obtain a basis U_i for S_i. If we knew a data point $\mathbf{x}$ for each subspace, we can recursively retrieve a basis U_i for each subspace S_i. See [104] for more details.

The correlation matrix of the Veronese embeddings in Eqn 4.20 defines the identifiable statistic $\Sigma_\mathcal{P}$ associated to the subspace clustering model[2] and we can therefore apply the compressive semi-parametric framework to it. As expected, the framework motivates us to seek structural assumptions of the statistic set $\mathcal{S}(\mathcal{X}) \subset \mathbb{R}^{D\times D}$. In the situation of GPCA, the correlation has rank R between 1 and D depending on the geometric makeup of the subspaces as discussed above. In certain cases, the rank of the correlation matrix is small and therefore the degrees of freedom are far less than the dimensions of the statistic set. In such situations, we know that only $\mathcal{O}(DR)$ measurements are needed to recover $\mathbf{C}_\nu$ and therefore it is sufficient to take $m = \mathcal{O}(DR)$ rank-one projections of $\mathbf{C}_\nu$ to enable stable recovery [108]. The compressive GPCA algorithm reduces to a low-rank matrix recovery algorithm as discussed in Section 2.3.1 and therefore can be defined by the encoding-decoding pair $(\Delta, \mathcal{A})$:

[2]The Veronese correlation matrix is not the only identifiable statistic of the subspace clustering problem. For example one could add the assumption that each of the subspaces are Gaussian distributed and therefore reduce the problem to a mixture of probabilistic PCA model. Here one could as easily use the ECF of the model as the identifiable statistic. However, this would reduce the problem to a restrictive parametric one. See [107] for more details.

$$\begin{cases} \mathcal{A} : \mathbf{C}_\nu \in \mathbb{R}^{D\times D} \mapsto (\text{trace}(\mathbf{a}_i\mathbf{a}_i^T\mathbf{C}_\nu))_{1\leq i\leq m} \in \mathbb{C}^m \\ \Delta(\mathbf{y}_N, \mathcal{A}) = \arg\min_{\mathbf{C}} \|\mathbf{y} - \mathcal{A}(\mathbf{C})\|_2 \qquad \text{s.t} \quad \|\mathbf{C}\|_* \leq R \end{cases} \tag{4.21}$$

4.4.2 Comparison

In this section, we present a short case study on when compressive semi-parametric learning succeeds and when it does not by comparing the compressive ICA scheme introduced in Chapter 3 and the compressive GPCA method that was introduced in Section 4.4.1. Both learning tasks are semi-parametric and follow the reformulated compressive learning framework set out in Section 4.3 due to their reliance on an identifiable statistic $\Sigma_{\mathcal{P}}$ to infer the parameters θ of the model. When does compressive semi-parametric learning actually work? To quantify this, we compare the size of the sketch m with the size of the full data matrix $\mathcal{O}(Nd)$ as the complexities associated with inference rely on these dimensions.

Recall from Chapter 3 that in the compressive ICA scheme, a sketch of size $\mathcal{O}(n^2)$ suffices to accurately infer the parameters of the ICA model where n is the number of independent components. For simplicity let's assume the scenario where $n = d$. Then compression is achieved if and only if $\mathcal{O}\left(d^2\right) \leq \mathcal{O}\left(Nd\right)$. Subsequently, compressive semi-parametric learning attains compression when

$$d \lesssim N. \tag{4.22}$$

It is straightforward to see that, in the vast majority of cases, the compressive ICA scheme leads to substantial compression.

Next we focus on the compressive GPCA scheme introduced in Section 4.4.1. The sketch is of size $\mathcal{O}\left(DR\right)$ which depends on the rank R of the correlation matrix $\Sigma_{\mathcal{P}} \in \mathbb{R}^{D\times D}$. We therefore require

$$DR \lesssim Nd \tag{4.23}$$

for compression to be attainable. To contextualise Eqn 4.23, let the rank $R = \alpha D$ for $0 < \alpha \leq 1$ and from Eqn 4.18 note that $D \leq d^n$. Then it can be seen that compression is achieved when

$$\alpha \lesssim Nd^{-2n+1}. \tag{4.24}$$

In comparison to Eqn 4.22, it is not as abundantly clear when a compressive GPCA scheme

would lead to compression with respect to using the full data. To illustrate in real terms when compression is possible, we compute the compression ratio $\frac{DR}{Nd}$ for a fixed signal length of $N = 100000$ and vary the feature dimension d and the number of subspaces n. Figure 4.5 demonstrates the regions of attainable compression for different values of rank $R = \alpha D$ for $\alpha = 0.05$ and $\alpha = 0.8$ which represent a low-rank and high-rank correlation matrix, respectively. The size of the sketch is set at $m = 2\alpha D^2$ as recommended in the low-rank matrix recovery literature [35, 109]. A compression ratio of $\frac{DR}{Nd} \geq 1$ indicates the sketch achieves no compression. Even for a moderately large data length of $N = 100000$, one can see that the compressive GPCA scheme achieves compression in comparison to the full data matrix $\mathbf{X}$ for only relatively modest sizes of n and d. Consequently, compressive semi-parametric learning does not always achieve reductions in complexity.

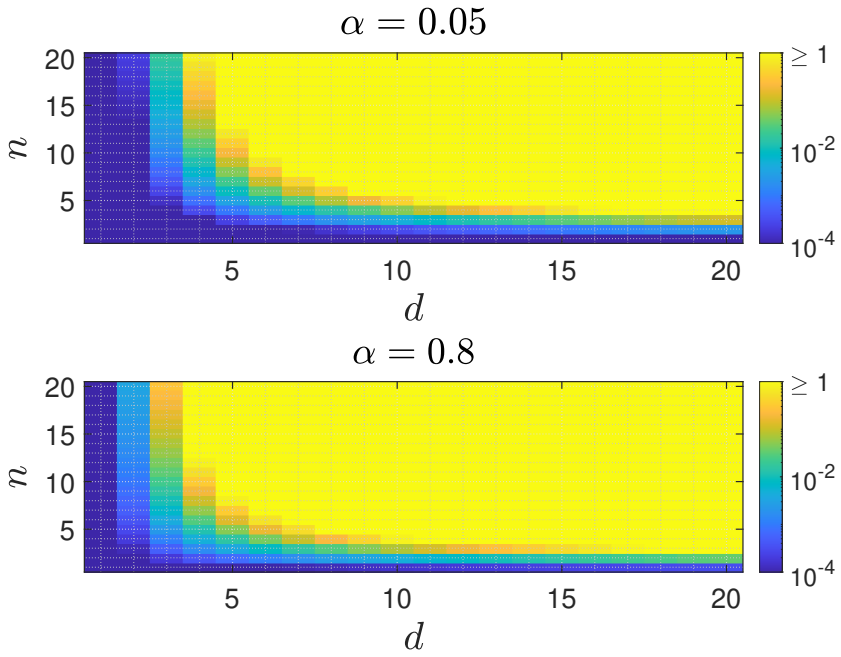

Figure 4.5: *The compression ratio $\frac{DR}{Nd}$ for compressive GPCA for a rank of $R = 0.05D$ (top) and $R = 0.8D$ (bottom).*

4.5 Discussion

In Section 4.4, we demonstrated that compressive semi-parametric learning does not always achieve compression, depending on the learning task. This is due in part to the reliance of an identifiable statistic that allows the parameters of the model to be estimated as presented

in the reformulated compressive learning framework in Section 4.3. In the case of GPCA, we leverage the correlation matrix of the Veronese embeddings $\mathbf{C}_\nu \in \mathbb{R}^{D\times D}$ where D has an exponential dependency on the model set's dimensions d and n. The regularity assumptions of the GPCA model set in Eqn 4.4 manifest structurally as a low-rank Veronese correlation matrix, therefore the sketch size exhibits a similar exponential dependency, i.e. $m = \mathcal{O}\left(\alpha D^2\right)$, and subsequently compression is only attained for modest sized model dimensions. In some cases, a semi-parametric model may not admit a finite dimensional identifiable statistic, which makes forming a compressive scheme very difficult.

On the other hand, we have observed successful semi-parametric learning schemes in the form of compressive PCA, as discussed in Section 2.3.3, and compressive ICA that was introduced in Chapter 3. The former method leverages the space of covariance matrices $\mathbf{C}_\mathbf{x} \in \mathbb{R}^{d\times d}$ to infer the PCA model parameters and requires only a sketch of size $m = \mathcal{O}(Kd)$ where K is the number of subspaces (i.e. the rank of the covariance matrix). The latter method utilizes the set of 4th order cumulant tensors $\mathscr{X} \in \mathfrak{C}$ and requires only a sketch of size $m = \mathcal{O}\left(n^2\right)$, In both instances, the sketch has only a quadratic dependency on the model's dimensions. Table 4.1 summarises the properties of the PCA, ICA and GPCA models that have been discussed in this thesis. In its current formulation, the success of a compressive semi-parametric model hinges on the existence of an identifiable statistic that has a structure and degrees of freedom that scale in a reasonable manner with respect to the model set dimensions. From a geometric perspective, we seek

$$\dim\left(\mathcal{S}\left(\mathcal{X}\right)\right) \;=\; \mathcal{O}\left(\dim\left(\mathfrak{S}_\theta\right)\right). \tag{4.25}$$

In other words, the dimension of the identifiable statistic space should be of the order of the dimension of the model set.

Semi-Parametric Compressive Learning Models				
Model	Stat. Hypothesis	Identifiable Statistic	Structural Redundancy	Sketch Size
PCA	Data lies on a K-dim. orthogonal subspace	Covariance matrix $\mathbf{C}_\mathbf{x}$	Low rank - K	$\mathcal{O}(Kd)$
ICA	Data is a mixture of n independent components	Cumulant tensor	Sparse tensor decomp.	$\mathcal{O}(n^2)$
GPCA	Data lies on a union of n subspaces of dim. d_i	Veronese Correlation matrix $\mathbf{C}_\nu$	Low rank - R	$\mathcal{O}(RD)$

Table 4.1: *A summary of the current semi-parametric compressive learning models and their properties.*

That being said, a compressive semi-parametric scheme exhibits some favourable advantages

compared to its parametric counterpart. The main one being that the identifiable statistic acts as an *intermediary* sketch in its own right by forming a map from an infinite dimensional probability space to a typically finite dimensional statistic space $\mathcal{S}(\mathcal{X})$ as depicted in Figure 4.6. Thereby, the compressive learning task reduces to a finite dimensional compressive sensing problem (see Section 2.3.1) where we can use a host of long-established tools to design tractable decoders and prove theoretical results, for example the restricted isometry property. Moreover, instead of designing intricate parametric sketching operators like RFFs which may require cross validation tuning, one can use very general randomized linear algebra techniques, for example subgaussian matrices or fast JL transforms, to build simple semi-parametric sketching operators.

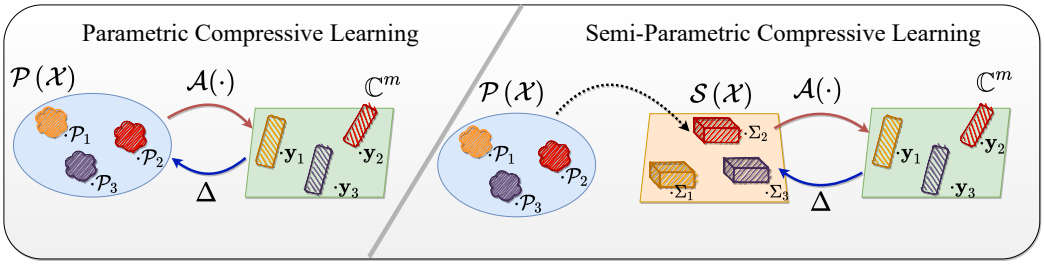

Figure 4.6: *A schematic diagram of parametric compressive learning (left) and semi-parametric compressive learning (right).*

4.6 Concluding Remarks

In this chapter, we proposed a reformulation of the original compressive learning framework that caters explicitly for the class of semi-parametric models. By leveraging an identifiable statistic associated with the semi-parametric model, we demonstrated that the compressive learning task reduces down to a finite compressive sensing problem where we instead attempt to sketch and then subsequently recover a statistic which has particular structure. We analysed the compressive subspace problem through the lens of the reformulated framework which highlighted that there may be instances where the summary statistic cannot always be effectively sketched. However there are also instances (e.g. compressive ICA) where the summary statistic can be effectively sketched, leading to significant compression. Future directions of research on compressive semi-parametric learning are discussed further in Chapter 7.

Part II

Part Two: Applications

Chapter 5
A Sketching Framework for Single Photon Counting Lidar

5.1 Introduction

Part I of this thesis concentrated on the theory and design of new compressive learning models specifically for the class of semi-parametric learning. Here in Part II, we focus on the applications of compressive learning, in particular, 3D depth imaging. As introduced in Section 2.5, single-photon counting lidar is an important tool in 3D depth imaging that can offer high temporal resolution over long range scenes. At the core of the technique is the ability to emit a pulse of light at a target using eye-safe lasers and then subsequently record the time-stamp of each individual photon that is detected by the single photon avalanche diode (SPAD). Over several clock cycles, a time correlated single photon counting (TCSPC) histogram is accumulated for each pixel in the field of view. An example of a TCSPC histogram for a given pixel is depicted in the left hand side of Figure 5.1 which counts the number of photons detected per histogram bin of time-interval $\Delta\tau$, where τ denotes the physical time-stamp over a given acquisition time. In general, a peak in the histogram implies the presence of a surface or object in the line of sight. Using the speed of light, we can simply convert the timing location of the peak(s) to determine the distance from the lidar device to the target.

In recent years, the rapid development of high rate, high resolution, low power ToF image sensors has caused a severe data transfer and processing bottleneck within the lidar device as a high resolution TCSPC histogram needs to be stored on chip or transferred off-chip for posterior depth estimation and other downstream tasks. As an example, take a high rate, high resolution lidar device capable of imaging a scene containing 256×256 pixels at a frame rate of 30 frames per second (fps). Then assuming the clock cycle of the laser (see Section 1.5) is discretized over $T = 2000$ time-stamp intervals where each TCSPC histogram is of 16 bit precision, then the lidar device would require a data transfer rate of 7.8 GB per second. Many techniques have been proposed to tackle the data transfer bottleneck of modern day lidar devices (see Section 2.5).

One of the most prominent of those is that of coarse binning which pools together adjacent time-stamp bins to form a coarse TCSPC histogram which requires less memory than the original TCSPC histogram (we delay formal definition until Section 5.3.3). The right hand side of Figure 5.1 shows an original TCSPC histogram coarsely binned into 50 timing bins. However, the compression achieved by coarse binning comes at the cost of losing the fine-grain temporal resolution provided by the original TCSPC histogram. Figure 5.2 demonstrates the loss of resolution incurred by estimating the depth and intensity parameters of the lidar observation model when estimated using the coarser histogram of 50 bins. As a consequence, the method of coarse binning forms a trade-off between compression and temporal resolution which is also inherent in the proposed schemes discussed in Section 2.5.

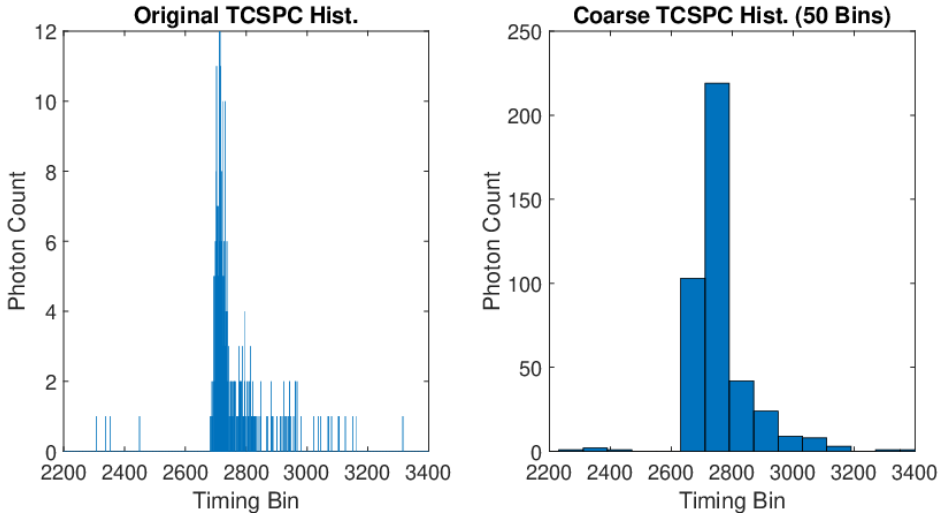

Figure 5.1: *An original zoomed in (around the target) TCSPC histogram with $T = 4613$ bins (left) and the coarse version of 50 bins (right)*

In this chapter, we propose a novel solution to this bottleneck of existing lidar techniques by calculating an on-the-fly sketch based on samples of the characteristic function of the ToF model. The size of the sketch scales with the degrees of freedom of the ToF model (i.e. number of objects in depth) and not with the number of photons or the fineness of the time resolution, without sacrificing precision in depth. The sketch can be computed for each incoming photon in an online fashion, only requiring a minimal amount of additional computation which can be performed efficiently on-chip. The sketch can be shown to capture all the salient information of the histogram, including the ability to explicitly remove background light or dark count effects, in a compact and data-efficient form, suitable for both on-chip processing or off-chip post processing. Furthermore, we develop a compressive lidar image reconstruction algorithm which

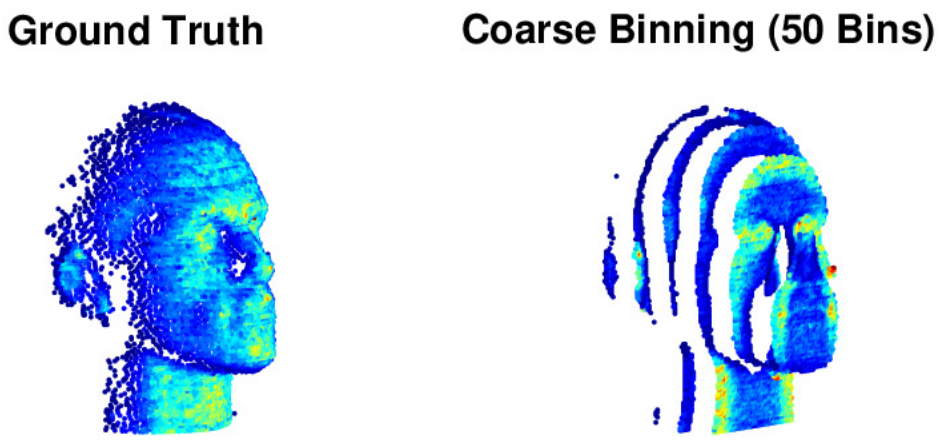

Figure 5.2: *The ground truth 3D depth image of a polystyrene head (left) and reconstruction using 50 coarse bins (right). The coarse binning method suffers from the staircase effect.*

has computational complexity dependent only on the size of the sketch. Our proposed method paves the way for high accuracy 3D imaging at fast frame rates with low power consumption. In summary the main contributions of the chapter are as follows:

- We propose a principled approach for compressing time-of-flight information in an online fashion without the requirement to form a histogram and without compromising depth resolution.
- A compressive single-photon lidar algorithm, named sketched maximum likelihood estimation (SMLE), is proposed which does not scale with either the number of photons or the time-stamp resolution in terms of space and time complexity.
- The statistical efficiency, given a compression rate (or sketch size), is quantified for different single-photon lidar scenarios, showing that only limited measurements of the characteristic function are needed to achieve negligible information loss.
- We analyse the performance of our proposed SMLE algorithm on both real and synthetic datasets and demonstrate the compression attained over other competitive techniques.

This chapter is based of the work in [110] that appeared in the IEEE Transactions on Computational Imaging. The remainder of the work is organized as follows. In Section 5.2 we detail the construction of the sketch using two different sampling schemes and we further demonstrate how our sketched lidar approach can be implemented in an online processing manner. In Section 5.3 we detail our proposed compressive single-photon reconstruction algorithm that

has computational complexity which scales with the sketch size m as well as quantifying the statistical efficiency of the estimated parameters θ. Results of the compressive lidar framework are analysed on both synthetic and real datasets in Section 5.4. Section 5.5 finally summarizes the chapter with some concluding remarks.

5.2 Sketched Lidar

We start this section with a warm up example to highlight the potential of using a sketch for single-photon lidar and to motivate the design of the sketch sampling procedure which will be discussed in Section 5.2.2. Before we do so, let's recall the lidar observation model from Section 2.5 where we assume there are K distinct reflecting surfaces within a given pixel. Denote by α_k and α_0 the probability that the detected photon originated from the kth surface and background sources, respectively and assume that for a single pixel, a total of N photons are detected during the whole acquisition window of the lidar device. Moreover, let τ denote the physical time-stamp such that the discretized time-stamp is denoted $t = \frac{\tau}{\Delta\tau}$. The time-stamp x_p of the pth photon where $1 \leq p \leq N$ can therefore be described by a mixture distribution [64]

$$\mathcal{P}(x_p|\alpha_0, ..., \alpha_K, t_1, ..., t_K) = \sum_{k=1}^{K} \alpha_k \mathcal{P}_s(x_p|t_k) + \alpha_0 \mathcal{P}_b(x_p), \tag{5.1}$$

where t_k denotes the discretized time-stamp of kth surface, α_k denotes probability that the detected photon originated from the kth surface and α_0 denotes the probability that the detected photon originated from background sources. In addition, $\sum_{k=0}^{K} \alpha_k = 1$. The distribution of the photons originating from the signal and background are defined by the distribution $\mathcal{P}_s(x_p|t) = h(x_p - t)/H$ and the uniform distribution $\mathcal{P}_b(x_p) = 1/T$ over $[0, 1, \ldots, T-1]$, respectively. Recall that h denotes the impulse response function of the system that can either be modelled or approximated via data-driven techniques and $H = \sum_{t=0}^{T-1} h(t)$ denotes the integral of the impulse response over the whole clock cycle.

5.2.1 Compressing Single Depth Data

In the absence of photons originating from background sources and the presence of a single surface or object, the sample mean of all the photon time-stamps ($\Phi(x) = x$) is the simplest summary statistic for estimating a single location parameter t_1. This only holds in the noiseless

case as the sample mean estimate is heavily biased toward the centre of the histogram when background photons are detected.

Suppose, we instead observe the cosine and sine of each photon count x with angular frequency $\omega = \frac{2\pi}{T}$, namely

$$\Phi(x) = \begin{bmatrix} \cos\left(\frac{2\pi x}{T}\right) \\ \sin\left(\frac{2\pi x}{T}\right) \end{bmatrix}, \tag{5.2}$$

and denote $\mathbf{y}_N$ the real valued sketch of size 2 ($m = 1$) computed over the dataset $\mathbf{X}$ as in Eqn 2.63. It is possible to recover an estimate of the single depth location parameter t_1 directly from the sketch via the trigonometric sample mean

$$\hat{t}_1 = \frac{T}{2\pi} \operatorname{phase}\left\{ \sum_{j=1}^{N} \cos\left(\frac{2\pi x_j}{T}\right) + \mathrm{i} \sum_{j=1}^{N} \sin\left(\frac{2\pi x_j}{T}\right) \right\} \tag{5.3}$$

where phase denotes the phasor angle. As the background photons are distributed uniformly over the interval $[0, T-1]$ ($\mathcal{P}_b(x) = \frac{1}{T}$), the expected moment of the photons originating from background sources is zero, $\mathbb{E}_{x\sim\mathcal{P}_b}\Phi(x) = \mathbf{0}$. The resulting estimate of the single depth parameter $\hat{t}_1$ is therefore an unbiased estimator (see Definition 2) of the location parameter t_1. The estimator in Eqn 5.3 coincides with the circular mean estimator detailed in [111]. Here the circular mean requires the first (non-zero) frequency.

We summarise the above using a simulated example, where a pixel of $T = 1000$ histogram bins with a detection point[1] signal-to-background ratio (SBR), defined as $\frac{1-\alpha_0}{\alpha_0}$, of 1 and a total of $N = 600$ photons is simulated, where the time-stamp of each photon is denoted by $\mathbf{X} = \{x_i\}_{i=1}^{N}$. The data was simulated using a Gaussian impulse response function with $\sigma = 15$ and a true position at time-stamp $t_1 = 320$. Computing the sketch $\mathbf{y}_N$ using the moment function from Eqn 5.2 and the associated circular mean estimate in Eqn 5.3, we obtain the sketch estimate $\hat{t}_{\text{cm}} = 323.3$ and the sample mean estimate of $\hat{t} = 434.1$. The TCSPC histogram along with both the circular and standard mean estimates as well as the location parameter t_1 are shown in Figure 5.3 where it is evident that the circular mean estimate does not suffer from the noise bias inherent in the sample mean.

Importantly, the sketch formed using the moment in Eqn 5.2 is equivalent to the complex valued

[1]Throughout the thesis, we consider the detection point SBR and not the raw sensor SBR which can be much lower in practice [112]

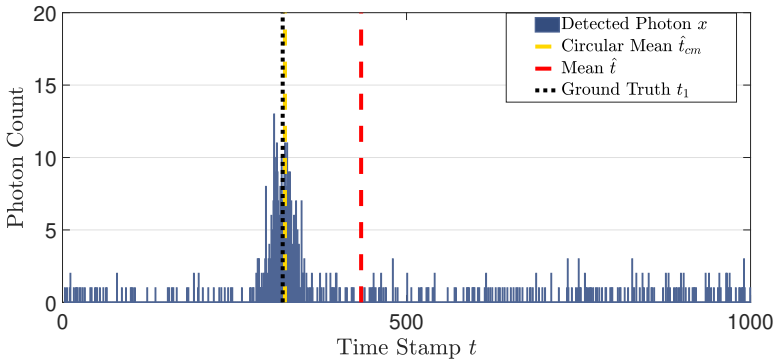

Figure 5.3: *The TCSPC histogram with $t_1 = 320$. The circular mean estimate (yellow) and the standard mean estimate (red) superimposed.*

ECF sketch (see Section 2.3.2.3)

$$\mathbf{y}_N = \frac{1}{N}\sum_{j=1}^{N} e^{\mathrm{i}\omega x_j} \in \mathbb{C}, \tag{5.4}$$

sampled at the point $\omega = \frac{2\pi}{T}$ and decoupled into both its real and imaginary components. In other words, we sample the empirical characteristic function (ECF) at the frequency $\omega = \frac{2\pi}{T}$. For the observation model defined in Eqn 5.1 and given a discrete impulse response function h, the characteristic function of the observation model is defined as

$$\begin{aligned} \Psi_{\mathcal{P}}(\omega) &= \sum_{k=1}^{K} \alpha_k \Psi_{\mathcal{P}_s}(\omega) + \alpha_0 \Psi_{\mathcal{P}_b}(\omega) \\ &= \sum_{k=1}^{K} \alpha_k \hat{h}(\omega) e^{\mathrm{i}\omega t_k} + \alpha_0 D_{\frac{T-1}{2}}(\omega) \end{aligned} \tag{5.5}$$

where $D_n(\omega) = \frac{\sin((n+1/2)\omega)}{2\pi \sin(\omega/2)}$ is the Dirichlet kernel function [113] and $\hat{h}$ denotes the (discrete) Fourier transform of the impulse response function h. Straightaway, we see why the choice $\omega = \frac{2\pi}{T}$ leads to an unbiased estimator: as $D_{\frac{T-1}{2}}(\omega) = 0$, the characteristic function $\Psi_{\mathcal{P}}$ is sampled at a point where the background photon's component (i.e. the noise) is zero. In fact, the estimate $\hat{t}_1$ in Eqn 5.3 is the optimal estimator to the compressive ECF sketch detailed in Eqn 2.16 (see Appendix 5.A).

Principally, we only need to store and transfer 2 values to accurately estimate the depth location of the object or surface, without the requirement to recourse to the original photon time-stamped

data. For the remainder of this section, we generalize the approach of forming a sketch of arbitrary size and sampling the ECF at multiple frequencies $[\omega_i]_{i=1}^m$. This will enable us to obtain statistically efficient estimates for the single surface case and to solve more complex lidar scenes including several surfaces with varying intensities where more salient information of the observation model is required.

5.2.2 Sampling the ECF

Recall that the observation model in Eqn 5.1 is discretized over the interval $[0, T-1]$ which we can consider to be a sufficient sampling if the distribution in Eqn 5.1 is approximately bandlimited. As a result, the characteristic function $\Psi_{\mathcal{P}}(\omega)$ has a finite basis characterized by the set of frequencies

$$\left\{ \frac{2\pi j}{T} \;\middle|\; j = 0, 1, \ldots, T-1 \right\}. \tag{5.6}$$

We can generalise the approach from Section 5.2.1 by sampling multiple frequencies from the finite basis in order to construct the ECF sketch. As is the case for the circular mean, the frequencies $\omega = \frac{2\pi j}{T}$ for $j = 1, 2, \ldots, T-1$ correspond to the zeros in the Dirichlet kernel function associated with the background pdf $\mathcal{P}_b$. We can therefore construct a sketch of arbitrary dimension m that is also *blind* to photons originating from background sources by avoiding the zero frequency $\omega = 0$ of the finite basis. As a result, we define the set of orthogonal frequencies by

$$\Omega \coloneqq \left\{ \omega_j = \frac{2\pi j}{T} \;\middle|\; j = 1, 2, \ldots, T-1 \right\}. \tag{5.7}$$

We coin this set the *orthogonal frequencies* as it defines regions over the interval of the observation model's characteristic function where the signal's contribution is orthogonal to the background's contribution.

In order to construct a sketch, we are ultimately interested in retaining sufficient salient information of the characteristic function $\Psi_{\mathcal{P}}$ such that we can identify and estimate the unique location and intensity parameters θ of the observation model $\mathcal{P}(x;\theta)$ defined in Eqn 5.1. It was discussed in Section 2.1.3.1 that the CF of a probability distribution decays in frequency, i.e. $\Psi_{\mathcal{P}}(\omega) \to 0$ as $\omega \to \infty$. Furthermore, as the observation model is discretized over the interval, we assume that the characteristic function of the observation model is approximately band-limited. A natural sampling scheme would therefore be to sample the first m frequencies

of the orthogonal frequencies Ω to capture the maximum energy of the CF. In other words, we could truncate the CF of the observation model whilst avoiding the zero frequency.

Alternatively, in [1, 114], provable guarantees for estimating mixture of Gaussian models have been provided, under certain conditions based on random sampling (see Section 2.3.1) of the CF. It is understood that the higher frequencies of the CF may provide further information to help discriminate distributions that are close in probability space. Moreover, if the CF decays slowly in frequency then the energy of the CF will be spread more throughout the set of orthogonal frequencies. We therefore provide an alternative sampling scheme whereby we randomly sample the set of orthogonal frequencies with respect to some sampling law Λ. In a similar design to the frequency sampling pattern proposed in [46], we sample the orthogonal frequencies by

$$(\omega_1, \omega_2, \ldots, \omega_m) \sim \Lambda_{\hat{h}}, \tag{5.8}$$

where $\Lambda_{\hat{h}} \propto \hat{h}$. To formalize, we consider the following sampling schemes in order to construct our ECF sketches:

1. Truncated Orthogonal Sampling: Sample the first m frequencies i.e. $j = 1, 2, \ldots, m$ from Ω.

2. Random Orthogonal Sampling: Sample the set of frequencies randomly, governed by the distributing law $\Lambda_{\hat{h}}$.

Depending on the circumstances of the lidar device we might expect one or the other sampling scheme to perform better.

5.2.3 Practical Hardware Considerations

As it was discussed in Section 2.3.4, one of the major advantages of forming a sketch $\mathbf{y}_N$ is that it is naturally amenable to online processing. Recall that for an arbitrary pixel in the scene, the resulting sketch that can be transferred off-chip is $\mathbf{y}_N = \frac{1}{N}\sum_{i=1}^{N} \Phi(x_i)$. Algorithm 4 demonstrates how the sketch for a given pixel is updated in real time during an acquisition window where N photons are detected by the SPAD array. For each photon arrival x_j during the acquisition window, an intermediate sketch is accumulated as well as an integer counter. Once the acquisition window is over, the resulting sketch is transferred off-chip for post-processing.

Algorithm 4 Sketch Online Processing

Initialisation: $\mathbf{y}_N = 0, N = 0$
while Acquisition Window **do**
  if New Photon Arrival x_j **then**
    $\mathbf{y}_N \leftarrow \mathbf{y}_N + \Phi(x_j)$
    $N \leftarrow N + 1$
  end if
end while
$\mathbf{y}_N \leftarrow \mathbf{y}_N / N$
Ensure: The sketch $\mathbf{y}_N$ is transferred off-chip for post-processing.

This is very beneficial as all that is needed to be stored on-chip is the sketch $\mathbf{y}_N$ of size $2m$ and an integer counter. As such, forming the sketch in an online processing manner, as in Algorithm 4, circumvents the need to compute and store a large histogram or store each individual photon time-stamp. Algorithm 4 is a similar to the NEWMA algorithm proposed by Keriven in [115] that efficiently computed the sketch in an online fashion. It should be noted that no further hardware is required to form the sketch and existing lidar devices can be easily adapted to implement our proposed technique.

The computation of the sketch itself requires the calculation of the Fourier features (i.e. $\cos(2\pi\omega_j/T)$ and $\sin(2\pi\omega_j/T)$) which would have to be computed in real time for each time-stamp. However, various efficient logic-based schemes already exist for performing such computations [116] based on either the classic CORDIC algorithms or polynomial approximations. Alternatively, in [117], Schellekens et al. show that in principle one can also replace the Fourier features by alternative periodic functions (e.g. square waves or triangle waves) in conjunction with random dithering. Subsequently, we will assume that we have access to sufficiently accurate sketch values for the remainder of the chapter and leave analysis of sketches constructed with limited fixed precision to Section 5.4.4.

5.3 Sketched Lidar Reconstruction

5.3.1 Statistical Estimation

Once the ECF sketch is constructed using either sampling scheme, we must estimate the parameters θ of the observation model $\mathcal{P}(x;\theta)$ solely from the sketch $\mathbf{y}_N$. In general, there is no closed form expression to estimate θ from the sketch of arbitrary size in contrast to the circular mean estimate in Eqn 5.3. It can be shown [118] that a complex valued ECF sketch $\mathbf{y}_N$ of size

m, computed over a finite dataset $\mathbf{X} = \{x_1, \ldots, x_N\}$, satisfies the central limit theorem (see Section 2.2.2). Formally, a sketch $\mathbf{y}_N \in \mathbb{C}^m$ converges asymptotically to a Gaussian random variable

$$\mathbf{y}_N \xrightarrow{\text{dist}} \mathcal{N}\big([\Psi_{\mathcal{P}}(\omega_j)]_{j=1}^m , N^{-1}\Sigma_\theta \big), \tag{5.9}$$

where $\Sigma_\theta \in \mathbb{C}^{m\times m}$ is a circulant matrix that has entries $(\Sigma_\theta)_{ij} = \Psi_{\mathcal{P}}(\omega_i - \omega_j) - \Psi_{\mathcal{P}}(\omega_i)\Psi_{\mathcal{P}}(-\omega_j)$ for $i, j = 1, 2, \ldots, m$. The asymptotic normality result in Eqn 5.9 naturally leads to a sketch maximum likelihood estimation (SMLE) algorithm that consists of minimising the following

$$\arg\min_\theta \ \frac{m}{2}\log\det(\Sigma_\theta) + N(\mathbf{y}_N - \mathbf{y}_\theta)^T\Sigma_\theta^{-1}(\mathbf{y}_N - \mathbf{y}_\theta), \tag{5.10}$$

where for convenience we denote $\mathbf{y}_\theta = [\Psi_{\mathcal{P}}(\omega_j)]_{j=1}^m$. For an observation model consisting of K surfaces and a general impulse response function h, recall that

$$\mathbf{y}_\theta = \left[\sum_{k=1}^K \alpha_k \hat{h}(\omega_j) e^{\mathrm{i}\omega_j t_k}\right]_{j=1}^m \tag{5.11}$$

and $\theta = (\alpha_0, \alpha_1, \ldots, \alpha_K, t_1, \ldots, t_K)$. Note that we have dropped the Dirichlet kernel function in Eqn 5.5 on the assumption that we are using one of the proposed sampling schemes. Minimising Eqn 5.10 is approximately equivalent[2] to minimising the compressive GeMM objective function

$$\hat{\theta} = \arg\min_\theta \|\mathbf{y}_N - \mathbb{E}_{x\sim\mathcal{P}}\Phi(x)\|_{\mathbf{W}}^2, \tag{5.12}$$

with the weighting function chosen to be $\mathbf{W} = \Sigma_\theta^{-1}$. The weighting matrix $\mathbf{W} = \Sigma_\theta^{-1}$ is asymptotically optimal in the sense that it minimises the variance of the estimator $\hat{\theta}$ from the sketch $\mathbf{y}_N$ [23] (see Section 2.1).

In practice Σ_θ is θ dependent as it is a function of the underlying parameters θ that are to be estimated. There are various well established methods in the GeMM and ECF literature [22, 23] that tackle the difficulty of approximating Σ_θ and estimating θ simultaneously. In [119], they use the K-L method which iteratively estimates Σ_θ and θ in a two stage procedure by fixing and updating one at a time. One can exploit the circulant structure of Σ_θ to compute the inversion $\mathbf{W} = \Sigma_\theta^{-1}$ in a computation complexity of $\mathcal{O}(m\log m)$. Some particular methods [120] fix Σ_θ after only a few iterations of the K-L approach to reduce the computational complexity of

[2]Note that we have dropped the $\log\det(\Sigma_\theta)$ term from Eqn 5.1 as in practice it has a negligible affect on the optimization landscape.

the algorithm, although this typically comes at the cost of introducing sample bias [121]. Occasionally, the covariance matrix is set throughout to be the identity, $\Sigma_\theta = I$, reducing Eqn 5.10 to a standard least squares optimization and a computational complexity of $\mathcal{O}(m)$, however this generally results in a less statistically efficient estimator $\hat{\theta}$ [22]. In this chapter, we estimate Σ_θ and θ simultaneously at each iteration. This approach is commonly referred to as Continuous Updating Estimator (CUE) [120] and obtains estimates that do not produce sample bias like the two-step K-L approach [121] and can often lead to more statistically efficient estimators [22]. However, the SMLE method is not restricted to the CUE and in certain situations practitioners may choose to sacrifice unbiased and efficiently optimal estimators for a reduced computational complexity by considering the other methods discussed.

The optimisation problem in Eqn 5.10 is also typically non convex and can suffer from spurious local minima. For the case when there is only a single surface, we initialise the SMLE algorithm using the analytic circular mean solution in Eqn 5.3 with minimal added computational overhead. From our experience with synthetic and real data, the circular mean estimate generally initialises the SMLE algorithm within the basin of the global minima, hence the issues associated with non-convex optimization are circumvented. For the case of multiple surfaces, we form a coarse uniform grid across $[0, T-1]^K$ and initialise at the smallest SMLE loss.

Remark 6. *In the orthogonal sampling scheme, one could alternatively zero-pad the sketch, perform an inverse FFT (iFFT) and find the maximum peak to estimate the depth position of the surface. However, this approach is fundamentally different to that of the orthogonal truncated sketch as the iFFT method is simply a low pass approximation of the TCSPC histogram whereas the proposed SMLE algorithm performs nonlinear parameter fitting. As a result, the iFFT method will be particularly inaccurate at distinguishing between closely spaced reflectors. In contrast to the proposed sketched lidar acquisition, the iFFT method does not take into account the particular nature of the IRF and achieves poor depth accuracy in the presence of a non-symmetric IRF (see Appendix 5.C). Furthermore, the iFFT approach requires $\mathcal{O}(T)$ off-chip memory complexity in comparison to $\mathcal{O}(m)$ of our proposed SMLE algorithm.*

5.3.2 Central Limit Theorem

One of the main advantages of the SMLE lidar approach from Eqn 5.9 is that even at low photon levels (i.e. small N), the SMLE estimates quickly follow the central limit theorem (CLT) (see Section 2.2.2) and provide a good approximation of its expectation. In contrast, the TCSPC

histogram used for many estimation methods, discussed in Section 2.5, is a poor approximation to its expectation as each time-stamp bin t has only a small number of photons. Thus efficient processing of the full histogram data requires careful consideration of the underlying Poisson statistics [122]. This is illustrated in Figure 5.4 which shows four separate histograms of the error $(\hat{t} - t_1)$ for increasing photon count N, along with the asymptotic Gaussian distribution from Eqn 5.9. The estimate $\hat{t}$ was obtained from a real valued sketch of size 2 ($m = 1$) using the circular mean estimate in Eqn 5.3. The simulated data was the same as the motivation example in Section 5.2.1 where a Gaussian IRF with $\sigma = 15$ was used. The SBR was set at 1 and the total number of time-stamps was $T = 1000$. The total photon count varied from $N = 10$ to $N = 10000$ increasing by a factor of 10 each time. For each photon count, we estimated the location parameter t_1 a total of 1000 times where the data $\mathbf{X} = \{x_i\}_{i=1}^{m}$ was simulated independently for each trial.

Even at extremely low photon counts of $N = 10$, the error $(\hat{t} - t_1)$ can be reasonably approximated by a Gaussian random variable centred around 0. This suggests that the estimate $\hat{t}$ quickly satisfies the CLT with respect to the photon count N. Further analysis of the proposed SMLE algorithm in the photon starved regime can be seen in Appendix 5.B. In the large photon regime ($N = 10000$), the estimation error is concentrated tightly around zero and mostly contained within 5 time-stamps. These results suggest that the sketched lidar CLT results of Eqn 5.9 hold even for low photons levels, hence the SMLE loss in Eqn 5.10 is a well-justified loss to minimise. A further potential benefit from this asymptotic normality is that it permits us to directly use *plug-and-play* Gaussian denoising algorithms to further improve the imaging performance [3, 123] which will be demonstrated in Chapter 6.

5.3.3 Statistical Efficiency

In this section, we calculate the theoretical statistical efficiency of the sketched lidar estimates, θ, that parametrize the observation model $\mathcal{P}(x; \theta)$ in Eqn 5.1, and compare them with the estimates obtained using the full data (i.e. no compression) using the relative error percentage. The relative error percentage, which will be defined later, is a key metric allowing us to quantify the relative loss of information given a sketch of size m from a statistical point of view. For sake of fair comparison to existing hardware implemented methods in the literature, the results and figures presented represent a sketch of size $2m$ where the real and imaginary components of the complex ECF sketch are stacked on top of each other to form an equivalent sketch consisting

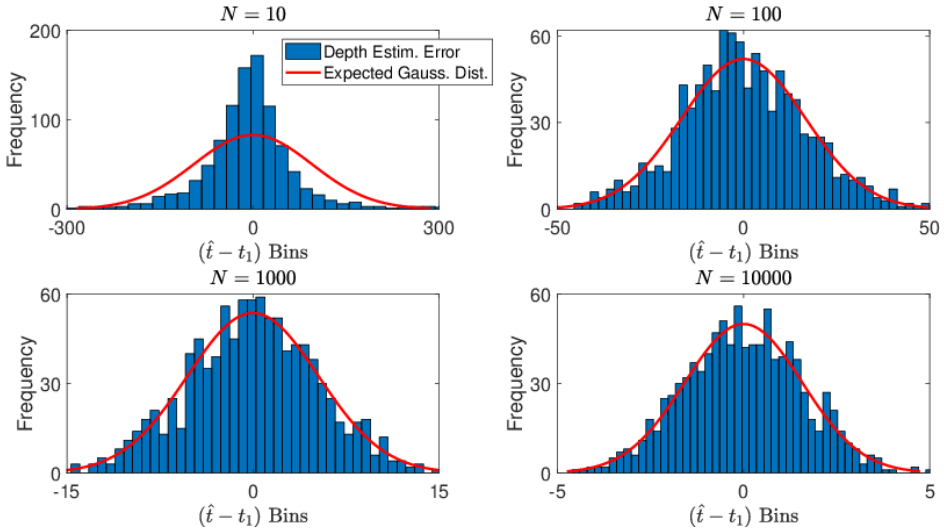

Figure 5.4: *Histograms of the estimation error $(\hat{t} - t_1)$ for increasing photon count N where the sketched lidar estimate (circular mean) is denoted by $\hat{t}$. The expected error distribution in Eqn 5.9 is depicted in red.*

of $2m$ real valued measurements.

It was discussed in Section 2.2.3 that the statistical efficiency is a measure of the variability or quality of an unbiased estimator $\hat{\theta}$ [20]. The Cramér-Rao bound gives a lower bound on the mean squared error of $\hat{\theta}$ [21] and therefore provides a best case scenario on the variability of the parameter estimates. Given the observation model $\mathcal{P}(x;\theta)$ and the corresponding Fisher information matrix (FIM), defined here as

$$\mathcal{I}_{\text{data}}(\theta) \coloneqq N\mathbb{E}\left[\left(\frac{\partial \log \mathcal{P}(x;\theta)}{\partial \theta}\right)^2\right], \tag{5.13}$$

then the optimal Cramér-Rao mean squared error, in terms of the full data, is defined as

$$\text{RMSE}_N \coloneqq \sqrt{\sum_{k=1}^{2K} [\mathcal{I}_{\text{data}}(\theta)^{-1}]_{\{kk\}}}. \tag{5.14}$$

Equivalently, we can compute the FIM for the sketched case using the normality result stated in Eqn 5.9, where the FIM of a multivariate Gaussian distribution [21] is defined as

$$\left(\mathcal{I}_{\text{sketch}}(\theta)\right)_{ij} \coloneqq N \frac{\partial \mathbf{y}_\theta}{\partial \theta_i} \Sigma_{\theta_0}^{-1} \frac{\partial \mathbf{y}_\theta}{\partial \theta_j}, \tag{5.15}$$

where $\mathbf{y}_\theta$ is the expected sketch defined in Eqn 5.11. Similarly, we define the optimal sketched

Cramér-Rao mean squared error as

$$\text{RMSE}_m \coloneqq \sqrt{\sum_{k=1}^{2K} [\mathcal{I}_{\text{sketch}}(\theta)^{-1}]_{\{kk\}}}. \tag{5.16}$$

To quantify the statistical efficiency of an estimate obtained from a real valued sketch of size $2m$, we use the relative error percentage (REP) metric which compares the optimal sketch root mean squared error RMSE_m with the corresponding full data root mean squared error RMSE_N, defined by

$$\text{REP} \coloneqq 100\left(\frac{\text{RMSE}_m - \text{RMSE}_N}{\text{RMSE}_N}\right). \tag{5.17}$$

Notably, the FIM of the sketched statistic in Eqn 5.15 scales with N, hence the REP metric is independent of the photon count. We compare the statistical efficiency of the sketched lidar estimates to the alternative compression technique of coarse binning [65] discussed at the start of this chapter in Section 5.1. The coarse binning approach can be seen to be equivalent to constructing a summary statistic

$$\tilde{\mathbf{z}}_n = \sum_{i=1}^{N} \left\{ \mathbb{1}_{[(j-1)\Delta_{\tilde{m}}, j\Delta_{\tilde{m}}]}(x_i) \right\}_{j=1}^{\tilde{m}}, \tag{5.18}$$

where $\Delta_{\tilde{m}} = \left\lceil \frac{T}{\tilde{m}} \right\rceil$ denotes the down-sampling factor, $\tilde{m}$ denotes the number of measurements equivalent to the real-valued sketch size (i.e. $\tilde{m} = 2m$) and $\mathbb{1}_{[t_i, t_i+\Delta_{\tilde{m}}]}(x)$ is the indicator function defined as

$$\mathbb{1}_{[(j-1)\Delta_{\tilde{m}}, j\Delta_{\tilde{m}}]}(x) \coloneqq \begin{cases} 1 & \text{if } x \in \big[(j-1)\Delta_{\tilde{m}}, j\Delta_{\tilde{m}}\big), \\ 0 & \text{Otherwise.} \end{cases} \tag{5.19}$$

Once the coarse binning sketch has been constructed, traditional estimation methods, for e.g cross correlation [124] or expectation maximization [10], can be employed to estimate the parameters of the observation model.

Lidar scenes typically have only 0, 1 or 2 reflectors in the scene, although in some specific applications, for example airborne lidar [125], tree-canopy foliage can return $K > 2$ reflectors. Our proposed method can handle greater number of reflections, however in the following experiments we only consider the typical case where $K = 1, 2$. Moreover, we choose the setting of the lidar scene (e.g. binning resolution, peak location, intensity) to best replicate a realistic

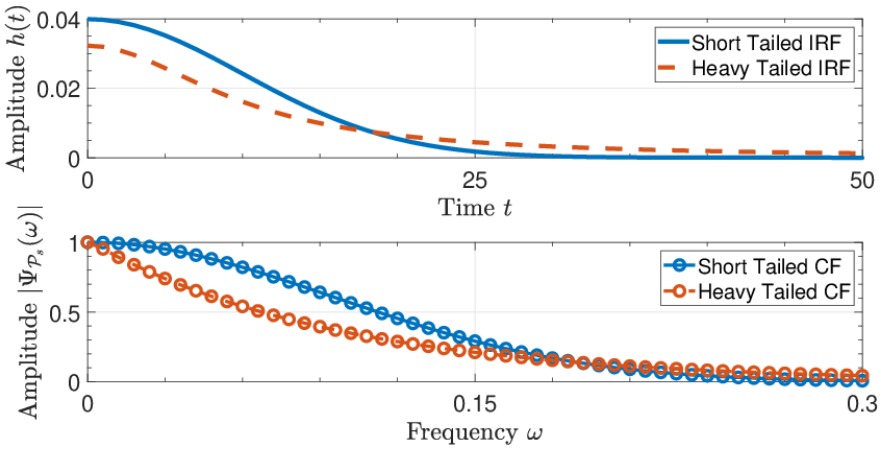

Figure 5.5: *The CF (top) of a short (blue solid) and long (red dashed) tailed impulse response function (bottom).*

setting as seen in Section 5.4.3. In each experiment, we consider two different impulse response functions (IRF), exhibiting both a short and long-tail. Figure 5.5 depicts the contrasting IRFs and the magnitude of their corresponding characteristic functions, $\Psi_{\mathcal{P}_s}(\omega) = \hat{h}(\omega)e^{\mathrm{i}\omega t}$. We evaluate the statistical efficiency of the sketched and coarse binning estimate using the REP as a function of the number of real measurements $2m$ and examine both the random and truncated orthogonal sampling schemes discussed in Section 5.2.2.

5.3.3.1 One Surface

We first evaluate the REP for a single peak case positioned at $t_1 = 430$, a window size of $T = 1000$. We consider both low and high background photon count levels, where the SBR was set at 10 and 1, respectively. Figure 5.6 shows the REP metric as a function of the number of real measurements $2m$ for the truncated orthogonal (blue), random orthogonal (red) and coarse binning (orange) compression techniques, where the high (SBR=10) and low (SBR=1) background photon levels are denoted by a solid and dashed line, respectively. The top and bottom plots depict the short and long-tailed IRF, accordingly. We first observe that both sketched lidar sampling schemes approach a REP of 0% as the real measurements increase and only a modest number of measurements is needed to obtain a low REP. In contrast, the coarse binning approach exhibits a slow convergence REP and remains high throughout the measurement range. Importantly, we see that the different sketch sampling schemes outperform each other depending on the tail of the IRF and hence the rate of decay of the CF. For instance, the truncated scheme produces a lower REP for the short-tailed IRF, while the random sampling scheme

achieves a quicker convergence and a significantly lower REP throughout the measurement range for the long-tailed IRF. This can be explained by Figure 5.5, the CF of the short-tailed IRF has the majority of its energy contained within the first few ($m = 10$) frequencies, while the CF of the long-tailed IRF has its energy spread more throughout its frequency.

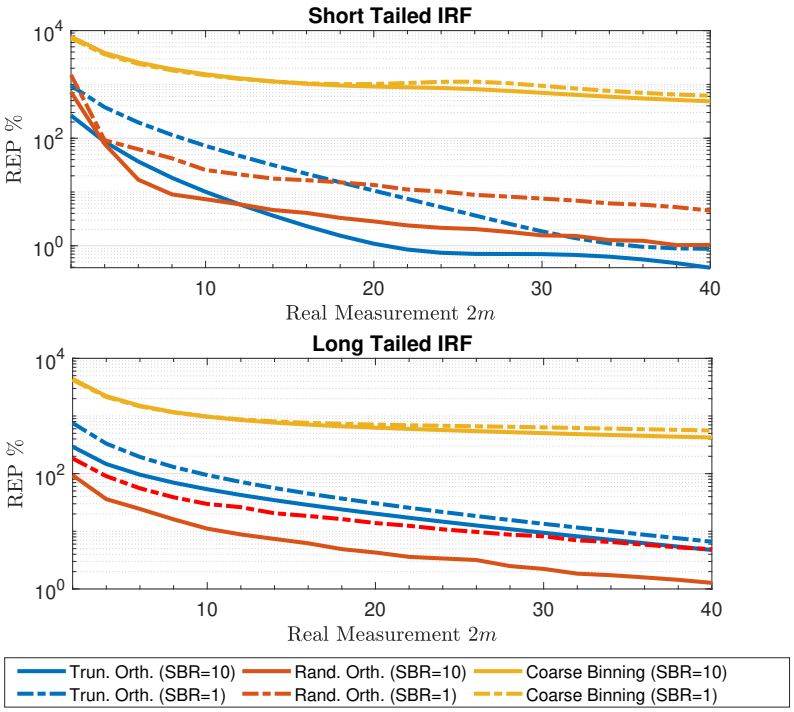

Figure 5.6: *The REP as a function of the number of real measurements ($2m$) for a single peak lidar scene.*

5.3.3.2 Two Surfaces

We now evaluate the REP for a two peak case positioned at $(t_1, t_2) = (320, 570)$, a window size of $T = 1000$. The intensity of the two peaks is given by 75% and 25%, respectively, simulating an object that is positioned behind a semi-transparent surface. We simulate both low and high background levels, where the SBR was again set at 10 and 1, respectively. Figure 5.7 shows the REP metric as a function of the number of real measurements $2m$ for the truncated orthogonal (blue), random orthogonal (red) and coarse binning (orange) compression techniques, where the high (SBR=10) and low (SBR=1) background photon levels are denoted by a solid and dashed line, respectively. The top and bottom plots depict the short and long-tailed IRF, accordingly. We see the same pattern as the single surface case where the REP remains high for the coarse binning compression technique while, in contrast, the sketched lidar converges towards a rel-

atively low REP in a modest number of measurements. We again observe that the truncated scheme performs best on a fast decaying CF, while the random sampling scheme outperforms the truncated counterpart when there is a slow decaying CF. The doubling of the dimension of the parameter θ by estimating two peaks and intensities, does not have a significant impact on the required number of measurements needed to achieve a relatively low REP. For instance in the high SBR (solid) scenario, the truncated orthogonal sampling scheme requires 20 real measurements ($m = 10$) to achieve a REP less than 1% for the unimodal case compared with a requirement of 24 real measurements ($m = 12$) to achieve the same level of REP for the bimodal case. These theoretical results on the statistical efficiency of the lidar sketch show that only a moderate sketch size is needed to achieve negligible loss of information. The results are based on the asymptotic normality property discussed in Eqn 5.9, and we have seen in Section 5.4.2 that in practice this normality result holds even for small photon counts of $N = 10$.

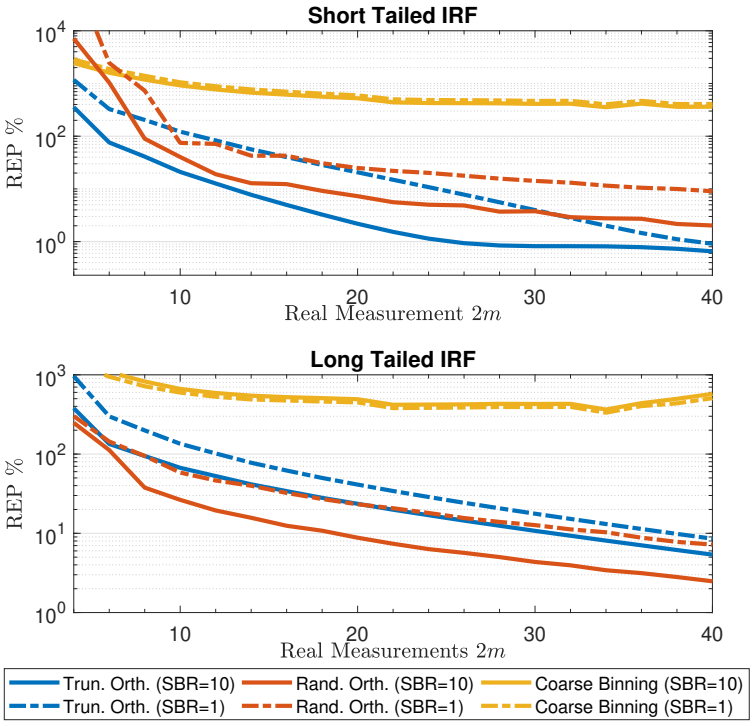

Figure 5.7: *The REP as a function of the number of real measurements ($2m$) for a lidar scene with 2 surfaces.*

In coarse binning, it can be beneficial to broaden the impulse response (while keeping laser power constant) such that it covers more than a single coarse bin. This strategy can achieve (coarse) sub-bin resolution (see for example [126]). Furthermore, Gyongy et al. [126] proposed

an algorithm that estimates the depth position continuously under the restricted assumption of a Gaussian IRF, in contrast to quantization limited cross correlation [124]. We further compare our proposed sketched lidar method to the wide pulse width coarse binning and algorithm used in [126] for a range of SBR values. In the simulation, the photon count was set at $N = 100$ and a Gaussian IRF was used. For the wide pulse width, we replicate the lidar device by setting $\sigma_1 = 0.4$. To compare with the narrow pulse width settings, we set $\sigma_2 = 5$. In both scenarios, a total of $2m = 16$ coarse bins are used. For our proposed SMLE algorithm, we compare the same compression be taking a (real-valued) sketch of size 16 ($m = 8$). We evaluate the depth estimation over SBR values ranging between 10^{-1} to 10^2 for 250 Monte-Carlo simulations. The coarse binning CRB is calculated where the best pulse width has been optimally selected for each SBR level.

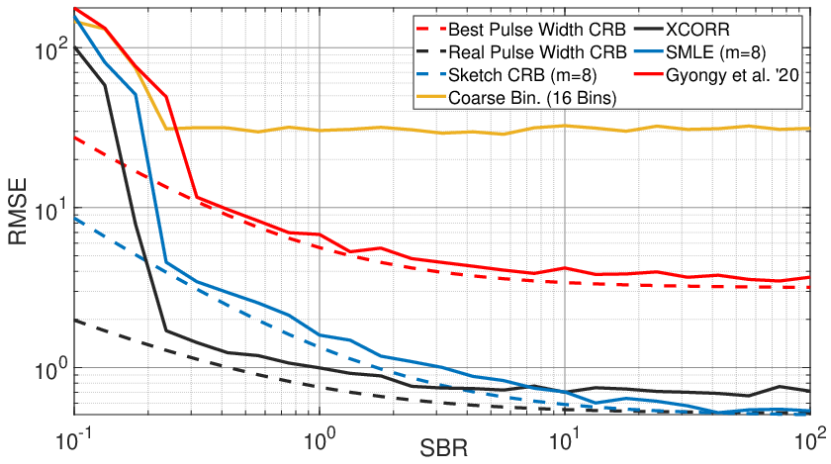

Figure 5.8: *Comparison of the RMSE achieved by wide and narrow Gaussian pulse width coarse binning to our proposed SMLE algorithm.*

As shown in Figure 5.8, the coarse sub-bin resolution can indeed improve the resolution with respect to coarse binning in large SBR regimes, but it still falls significantly behind the resolution obtained using the narrowest IRF with a fine scale time-stamp of our proposed sketch method. For instance, at an SBR of 0.23 the wide pulse width achieves a RMSE of 264.6 bins compared to 31.1 and 4.5 bins for the narrow pulse width coarse binning and SMLE, respectively. As the pulse width optimised algorithm in [126] only exhibits significant improvement in the high SBR scenario and is restricted to a Gaussian IRF, we do not consider it further in the chapter.

5.4 Experiments

5.4.1 Experimental set up

In this section, we evaluate our compressive lidar framework on synthetic and real data with increasingly complex scenes. Our method is compared with classical algorithms working on the full data space (i.e. no compression) namely cross correlation [124] and expectation maximization (EM) [10]. Moreover, we also compare our results to the alternative compression technique of coarse binning [65] discussed in Section 5.1 and Eqn 5.18. Both the cross correlation and EM algorithms estimate the location parameters using the full data and therefore the results obtained from these methods set a benchmark to the estimation accuracy when no compression takes place. For sake of fair comparison, we use the real valued sketch in all the subsequent results, such that the number of real measurements is equivalent to $2m$.

5.4.1.1 Processing

Restoration of depth imaging of single-photon lidar consists of estimating a 3D point cloud from a lidar data cube containing the number of photons $N_{i,j,t}$ in pixel (i, j) at time-stamp t, where $i = 1, 2, \ldots, N_r, j = 1, 2 \ldots, N_c$ and $t = 0, 1, \ldots, T-1$. We denote the average photon count for each pixel by $\bar{N}$ and process each pixel (i, j) of the data cube and estimate the true location and intensity parameter, denoted t_1 and α, respectively. The intensity of a point in pixel (i, j) of the point cloud is calculated by the number of photons in the pixel multiplied by the proportion of the signal i.e. $\alpha_k \sum_{t=0}^{T-1} N_{i,j,t}$. A data driven impulse response is given for each dataset and we can obtain the characteristic function of the IRF by using Eqn 5.5.

5.4.1.2 Evaluation Metrics

Two different error metrics are used to evaluate the performance of our proposed sketched lidar framework. We consider the root mean squared error (RMSE) between the reconstructed image and the ground truth. Given that $t_{i,j,k}$ is the location of the kth peak in pixel (i, j) and $\hat{t}_{i,j,k}$ the estimated counterpart, then the root mean squared error of the reconstructed image is

$$\text{RMSE} := \sqrt{\frac{1}{KN_rN_c} \sum_{i=1}^{N_r} \sum_{j=1}^{N_c} \sum_{k=1}^{K} \left(t_{i,j,k} - \hat{t}_{i,j,k} \right)^2}. \tag{5.20}$$

Moreover, we also use the mean absolute error metric defined as

$$\text{MAE} := \frac{1}{KN_rN_c}\sum_{i=1}^{N_r}\sum_{j=1}^{N_c}\sum_{k=1}^{K}|t_{i,j,k} - \hat{t}_{i,j,k}|. \tag{5.21}$$

The compression of both the sketched lidar and coarse binning approach is measured in terms of the dimension reduction achieved by the statistic with respect to the raw TCSPC data and is quantified by the metric $\max\{\frac{2m}{T}, \frac{2m}{N}\}$, which is dependent on the dimensions, T and N, of the lidar scene and where the number of real measurements $(2m)$ is used for sake of fair comparison.

5.4.2 Synthetic Data

We evaluate the sketched lidar framework on a synthetic dataset simulating a pixel in a scene which consists of a single peak response. We chose the parameters that best replicated a realistic lidar scene and that were akin to the real datasets which will be discussed in later in Section 5.4.3. Therefore, we set the binning resolution at $T = 250$, and impulse response was generated with a true Gaussian function where $\sigma = 5$. We ran a Monte-Carlo simulation with 1000 trials to evaluate and compare the performance of our sketched lidar framework for photon counts $N \in (100, 1000)$ with varying SBR levels and number of real measurements $2m$. For each trial, we uniformly chose $t_1 \sim \mathcal{U}(0, 249)$, and estimated $\hat{t}$ for the sketched lidar approach, the iFFT method discussed in Section 5.3 as well the alternative compression technique of coarse binning. As a reference, we computed the cross correlation estimate as well as estimating the maximum peak of the full histogram which represent the estimates over the full data (i.e. no compression). We varied the total number of real measurements between 2 ($m = 1$) and 50 ($m = 25$) and increased the SBR ratio from 10^{-2} to 10^2 on a log-scale. Here we only show the results for the truncated orthogonal sampling scheme but we observed in practice that the alternative random orthogonal sampling scheme produces similar results.

Figures 5.9 and 5.10 show the contour plots of the RMSE level of $10\Delta\tau$ (left) and $2\Delta\tau$ (right) (i.e. 10 and 2 time-intervals) for both $N = 100$ and $N = 1000$, respectively. The sketched lidar (solid blue), coarse binning (orange) and the iFFT (red) methods are depicted alongside the full data approaches of cross correlation (XCORR) (solid black) and maximum peak estimation (green). As discussed in Section 5.3.3, the full data (dashed black) and the sketched (dashed blue) Cramér-Rao bound are given as reference and define the lower bound to the contour plot.

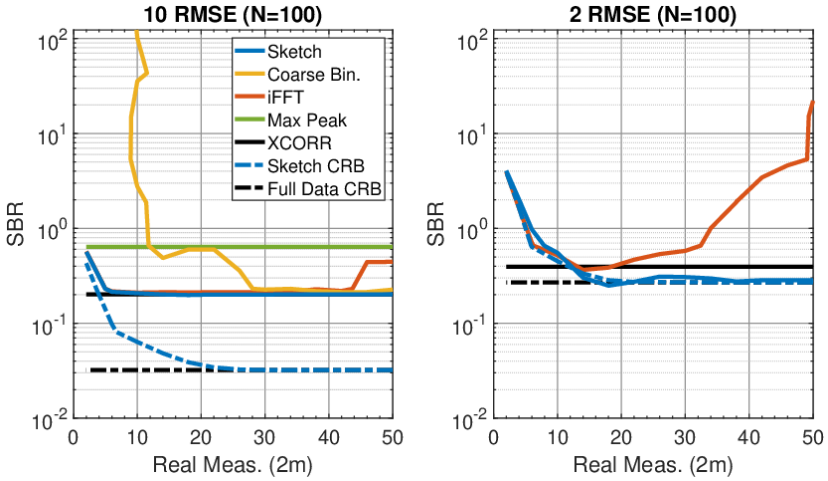

Figure 5.9: *RMSE level set contour plots for varying SBR levels and number of real measurements* $2m$ *for a photon count of* $N = 100$*. The RMSE level are* $10\Delta\tau$ *(left) and* $2\Delta\tau$ *(right). The legend is defined for both plots.*

Both the sketched lidar and iFFT approach converge quickly towards the full data estimate of cross correlation within 10 real measurements for both RMSE level sets and photon counts. In contrast, the coarse binning approach needs approximately 30 real measurements to achieve a similar performance as our sketched lidar method in achieving a RMSE of $10\Delta\tau$. Moreover, coarse binning does not attain an RMSE of $2\Delta\tau$ for $2m \leq 50$ hence does not appear in the right subplot of Figure 5.9. It can be seen for a larger number of real measurements, the iFFT approach begins to diverge. This is because for larger number of measurements, the iFFT produces a less smooth linear approximation of the histogram and therefore it is more challenging to estimate the depth position.

Figures 5.11 and 5.12 show the 95% of peaks detected within the level sets of $10\Delta\tau$ (left) and $3\Delta\tau$ (right). Our proposed sketch method achieves the same estimation performance as the full data XCORR approach within approximately 12 real measurements ($m = 6$) for all varying SBR ratios and photon counts. In contrast, the coarse binning approach requires approximately 45 real measurements, equating to a modest compression of 0.25, to achieve 95% of detections within $10\Delta\tau$. Furthermore, the coarse binning method could not achieve 95% of detections within $3\Delta\tau$ for all the real measurements considered. These initial results on synthetic lidar data for a range of different SBR ratios and photon counts highlight the clear trade-off between compression and loss of temporal resolution for the coarse binning approach. In contrast, our proposed sketched lidar method overcomes the trade-off between compression and loss of resolution and only requires a very modest sketch size to achieve the same estimation performance

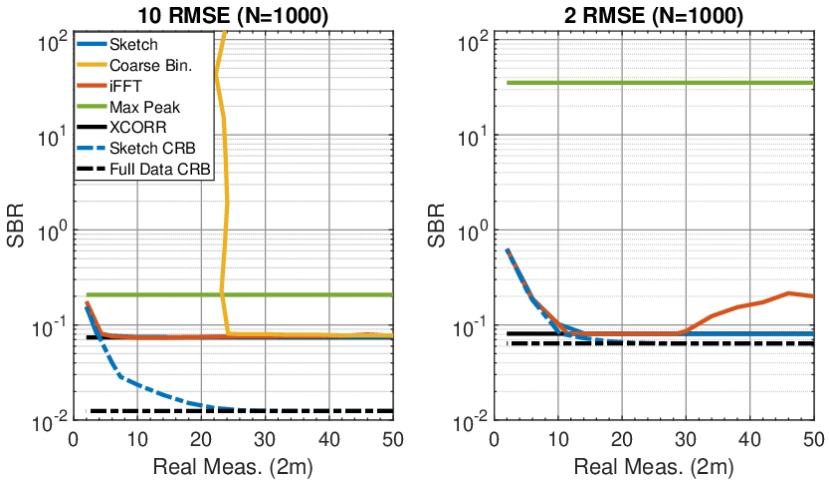

Figure 5.10: *RMSE level set contour plots for varying SBR levels and number of real measurements* $2m$ *for a photon count of* $N = 1000$*. The RMSE level are* $10\Delta\tau$ *(left) and* $2\Delta\tau$ *(right). The legend is defined for both plots.*

as cross correlation using the whole data.

5.4.3 Real Data

In this section we evaluate our sketched lidar framework on two real datasets of increasing complexity. Namely, a polystyrene head imaged at Heriot-Watt University [63, 127] which consists mostly of a single peak, and a scene where two humans are standing behind a camouflage net, depicted in [128, 3], which contains of 2 objects per pixel with varying intensity.

5.4.3.1 Polystyrene Head

The first scene consists of a polystyrene head placed 40 meters away from the lidar device. The data cube has width and height of 141 pixels, $N_r = N_c = 141$ and a total of $T = 4613$ time-stamps. A total acquisition time of 100 milliseconds was used for each pixel resulting in an average photon count of $\bar{N} = 337$ with an SBR of approximately 6.82. The vast majority of pixels consist of a single peak, although there are a minority of pixels around the borders of the head that consist of two peaks. The parameter set to be estimated for each pixel is $\theta = (t, \alpha)$ of dimension 2. We compare our results with the ground truth obtained from the experiment as well as the full data algorithm of XCORR and the coarse binning compression technique. As XCORR uses the maximum likelihood estimation of a single peak, we assume each pixel has one surface for the sake of comparison. As a result, we set the SMLE algorithm to estimate

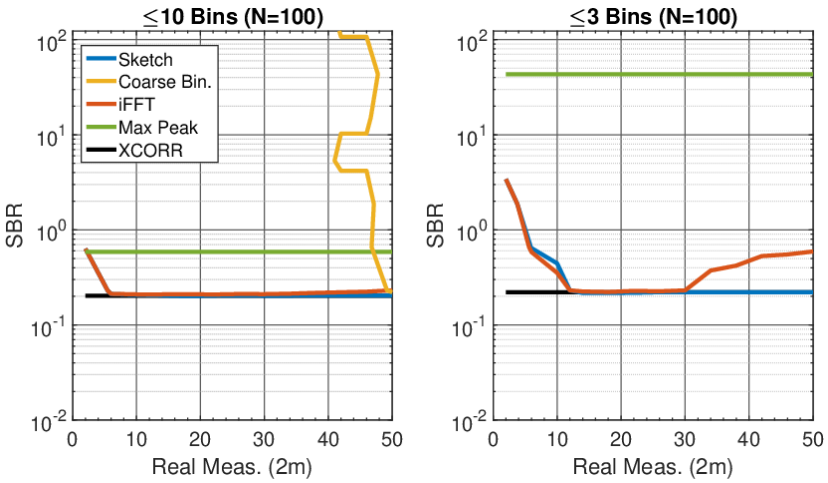

Figure 5.11: *RMSE level set contour plots for varying SBR levels and number of real measurements* $2m$ *for a photon count of* $N = 100$ *for detecting* 95% *of peaks within the level sets of* $10\Delta\tau$ *(left) and* $3\Delta\tau$ *(right). The legend is defined for both plots.*

a single peak, however in practice we can use detection algorithms, for instance the sketch-based detection scheme proposed in Chapter 6, to detect the number of surfaces present before estimation. The coarse binning approach is computed using cross correlation once the data cube is down-sampled.

The data driven impulse response function and its corresponding CF obtained from Eqn 5.5, are shown in Figure 5.13. We only present the results for the truncated orthogonal sampling scheme, from Section 5.2.2, but we observed in practice that the alternative random orthogonal sampling scheme produces similar results. We initialise the sketched lidar algorithm using the analytic circular mean solution in Eqn 5.3.

Figure 5.14 shows the reconstructed images of the sketched lidar, coarse binning and XCORR approaches, as well as the ground truth image. We first notice that our sketched lidar method sufficiently reconstructs the polystyrene head scene for all sketch sizes, even for the circular mean estimate ($m = 1$) in (a). In contrast, the coarse binning approach fails for all corresponding measurements $\tilde{m}$ with significant staircase artifacts arising. Figure 5.15 shows the RMSE, in comparison to the ground truth, as a function of the number of real measurements ($2m$). Here we omit the small proportion of pixels that consist of two peaks from the RMSE calculation for sake of fair comparison with the existing methods that can only estimate a single peak. We observe that our sketched lidar method produces a smaller RMSE as the measurement size increases and achieves a smaller RMSE than the LMF approach for larger measurements. In

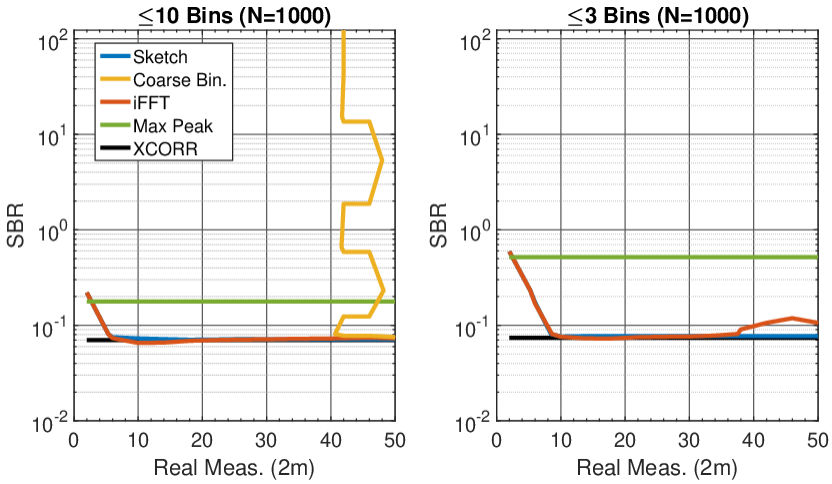

Figure 5.12: *RMSE level set contour plots for varying SBR levels and number of real measurements* $2m$ *for a photon count of* $N = 100$ *for detecting* 95% *of peaks within the level sets of* $10\Delta\tau$ *(left) and* $3\Delta\tau$ *(right). The legend is defined for both plots.*

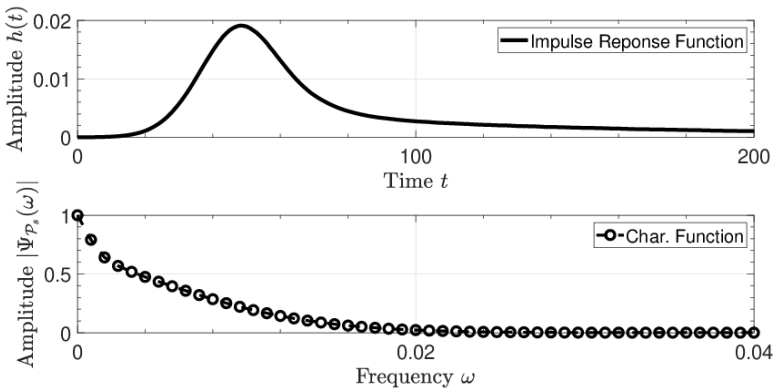

Figure 5.13: *The CF (bottom) of the data driven impulse response function (top) of the polystyrene head dataset.*

comparison, the coarse binning method obtain estimates that produce a large RMSE consistently throughout. As such, this suggests that our sketched lidar approach does not compromise reduced resolution in favour of compression which is very apparent in the coarse binning method.

5.4.3.2 Humans Behind Camouflage

The second scene consists of two humans standing behind a camouflage net approximately 320 metres away from the lidar device. Further details can be found of the scene in [128, 129]. The data cube has width and height of 32 pixels, $N_r = N_c = 32$ and a total of $T = 153$

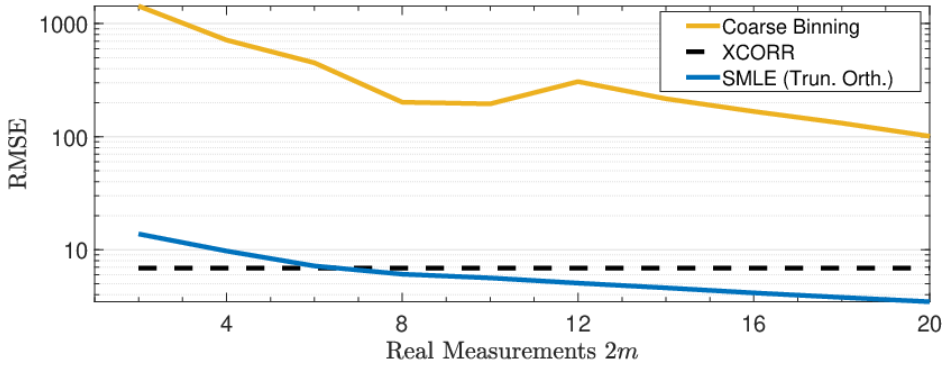

Figure 5.15: *The RMSE as a function of the number of real measurements ($2m$) for the polystyrene head dataset.*

time-stamps. A total acquisition time of 5.6 milliseconds was used for each pixel resulting in an average photon count of $\bar{N} = 871$ with an approximate SBR of 2.35. The vast majority of pixels have 2 surfaces (the camouflage net and a human) where the net (first peak) accounts for the biggest intensity. The parameter set to be estimated for each pixel is $\theta = (t_1, t_2, \alpha_1, \alpha_2)$ of dimension 4. We compare our results with the full data EM algorithm as well as the coarse binning compression technique. For this experiment, the coarse binning algorithm uses the EM estimate once the data cube has been down-sampled as the cross correlation algorithm is only applicable to single peak cases. Due to the lack of a ground truth, we compare the reconstructions of the camouflage scene to the full data EM algorithm reconstruction and equate the relevant compression of both the sketched lidar framework and the coarse binning technique. The data driven impulse response function h and its corresponding CF obtained from Eqn 5.5, are shown in Figure 5.16. Again, we only present the results for the truncated orthogonal sampling scheme, from Section 5.2.2, but we observed in practice that the alternative random orthogonal sampling scheme produces similar results. We uniformly sampled 10 starting points for each of peak t_1 and t_2 and initialised with the smallest sketched cost value from Eqn 5.10.

Figure 5.17 shows the reconstructed images of the sketched lidar, coarse binning and EM algorithm methods. Evidently, the reconstruction of our sketched lidar approach becomes better as the number of real measurements ($2m$) increases, for instance the torso of the human positioned near 600 cm has greater clarity in sketch size 20 compared to sketch size 4 where more spurious peaks are detected. However, the sketched lidar reconstruction for $m = 2$ is still sufficient in comparison to the EM reconstruction in (g), while in contrast the coarse binning method fails to reconstruct either human for the corresponding number of measurements. The coarse binning method once again suffers from the stair case effect as seen by the lack of width

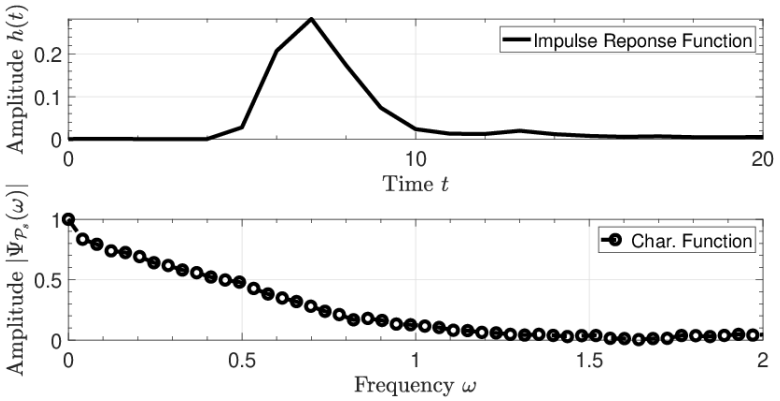

Figure 5.16: *The CF (bottom) of the data driven impulse response function (top) of the camouflage dataset.*

of the first human standing at position 200 cm in (f). Furthermore, the compression due to the coarse binning results in poor depth accuracy as seen by the position of the camouflage net in reconstruction (b) which has a disparity of approximately 120 cm in comparison to the EM reconstruction. Once again, this suggests that our sketched lidar approach does not compromise reduced resolution in favour of compression which is apparent in the coarse binning method.

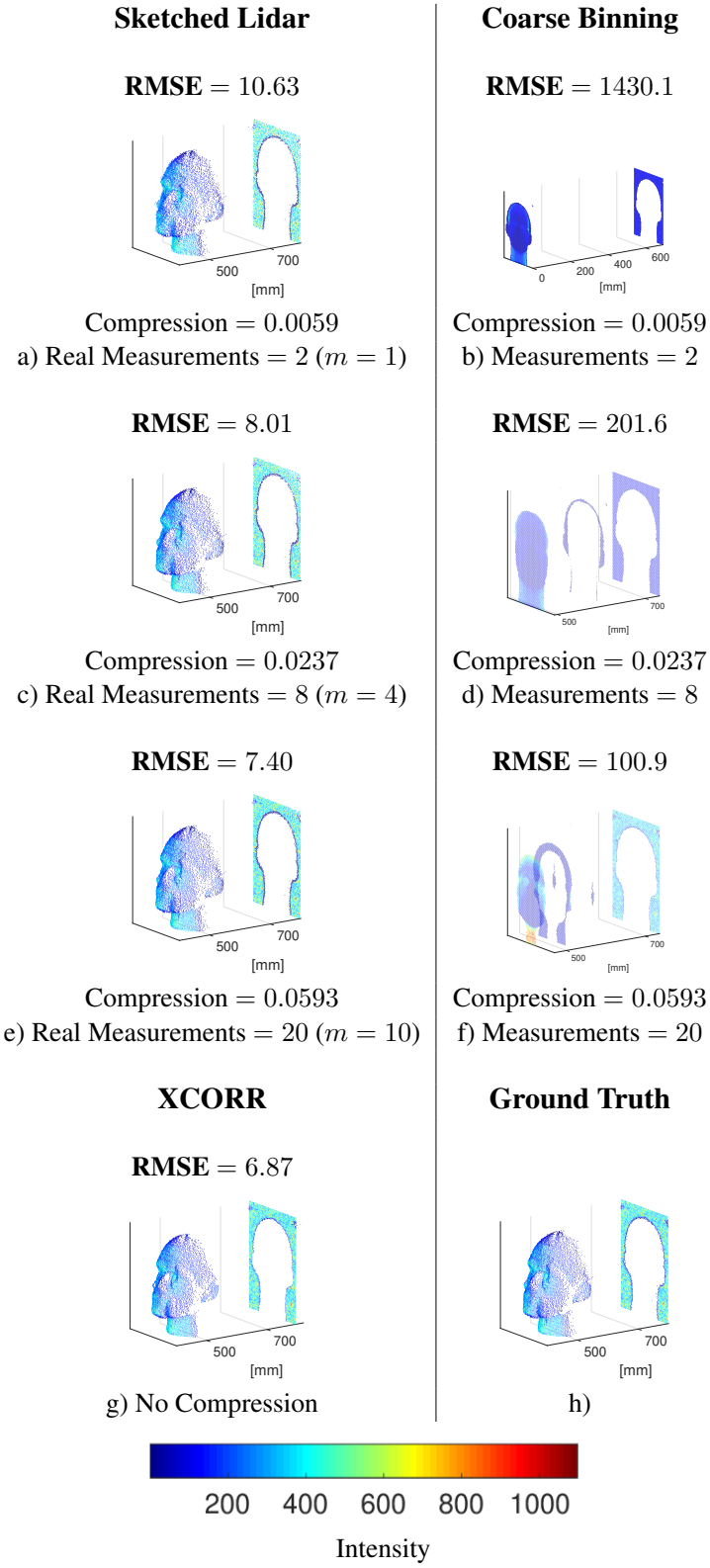

Figure 5.14: *The face dataset lidar reconstructions of the sketched lidar and coarse binning method for the real valued measurement size* $2, 8, 20$. *Both the cross correlation (XCORR) reconstruction and the ground truth image are given for comparison.*

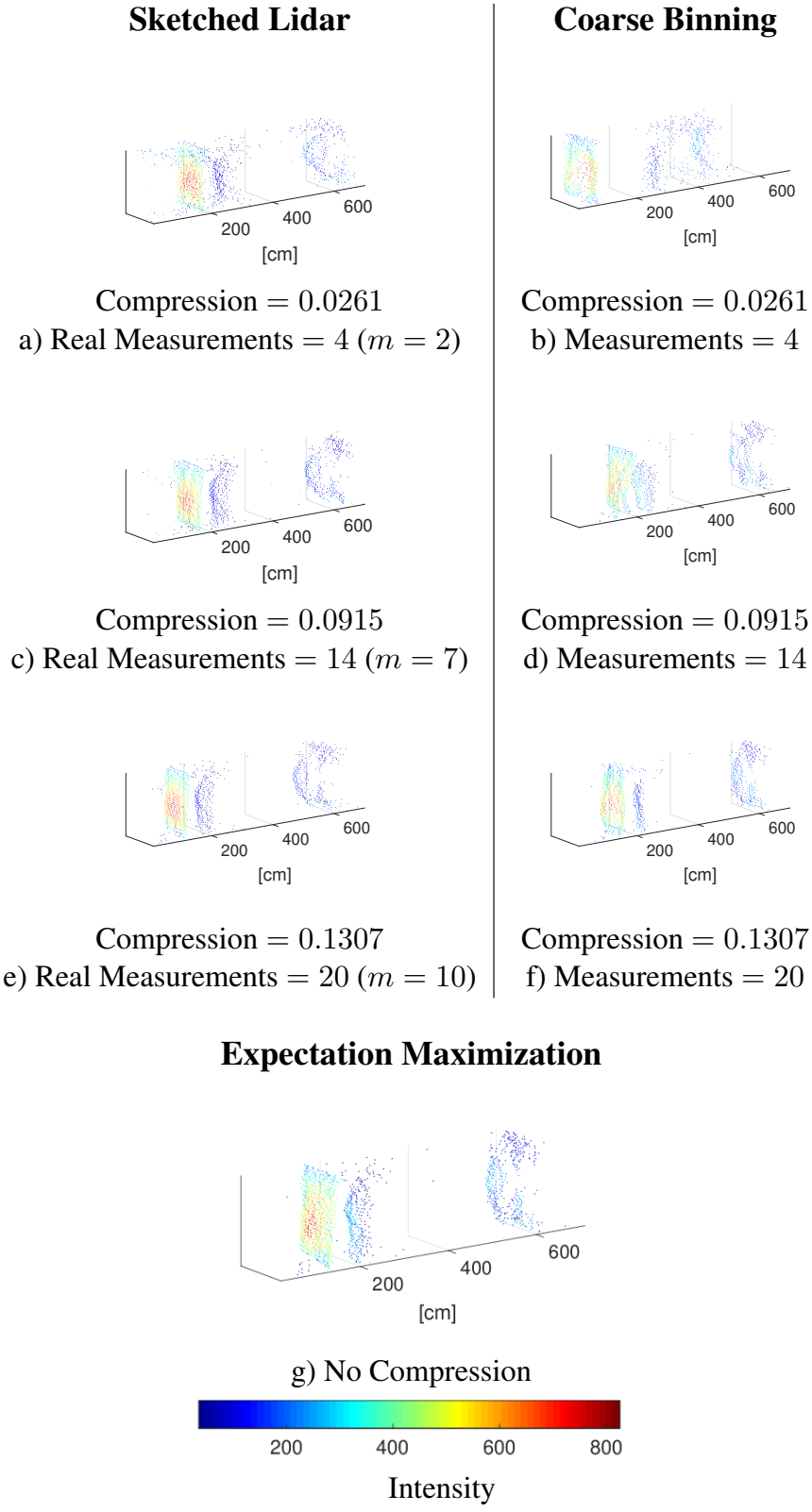

Figure 5.17: *The camouflage dataset lidar reconstructions of the sketched lidar and coarse binning method for the real valued measurement size* $(2m)$ *of* $2, 8, 20$. *Both the cross correlation (XCORR) reconstruction and the ground truth image are given for comparison.*

5.4.4 Wordlength Considerations

In Section 5.4.3, each individual entry of the sketch has a standard 32 bit precision. In practice, however, the logic within the lidar device (see Section 2.5) that constructs the sketches may constrain the level of precision one can work with. Here we analyse the quality of reconstruction depending on the the level of fixed precision for each individual entry of the sketch. We consider 32 (standard), 16, 12, 8 and 4 bit precision that is signed as the sketch values can be both positive and negative. Denoting by b the level of precision, we divide the bit budget into 1 sign bit, $b/2$ integer bits and $b/2 - 1$ fractional bits.

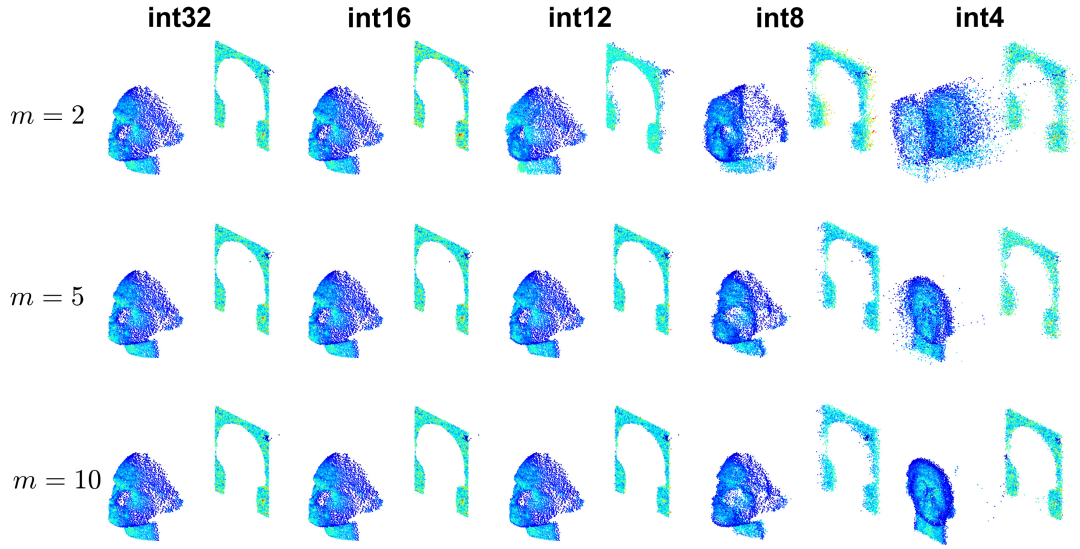

Figure 5.18: *The reconstructions of the polystyrene head dataset for a sketch of size* $m = 2, 5$ *and* 10 *where each sketch entry has a precision of 32, 16, 12, 8 and 4 bits.*

Figure 5.18 shows the reconstructions for a sketch of size $m = 2, 5$ and 10 where each sketch value has either a 32, 16, 12, 8 or 4 bit precision. Furthermore, Figure 5.19 shows the mean absolute error (MAE) of each sketch size for the associated entry-wise bit precision. Additionally, Figure 5.20 shows the total wordlength of the whole sketch such that we can compare like for like wordlengths. The results indicate that the the reduction of fixed precision effects smaller sized sketches to a larger extent. Notably, for the 12 bit precision reconstructions, the sketch of size $m = 2$ achieves a far worse reconstruction compared to its $m = 5$ and $m = 10$ counterparts. Interestingly, a sketch of size $m = 2$ with 32 bit precision achieves nearly the same MAE as a sketch of size $m = 10$ with 12 bit precision. However, for smaller SBR and photon count

values, a sketch of size $m = 2$ with 32 bit precision may incur a larger reconstruction error. Overall, the results show that for a sketch of size greater than $m = 2$, one can safely work with a reduced 12 bit precision without incurring significant loss of reconstruction quality.

In this section, we only consider the quantization of the overall bits used in the original sketch of Eqn 5.11. However, in [117] Schellekens et al. consider an asymmetric approach where the sketch feature function is different at the sketching phase to the learning phase to accommodate more hardware efficient sketch implementations. In this scenario, an additional random dither is added to the data before the sketch phase. It is shown that the additional dither ensures the asymmetric CL optimization procedure is not impacted by the quantization. We leave the implementation of additional dither to the sketched lidar framework for future work.

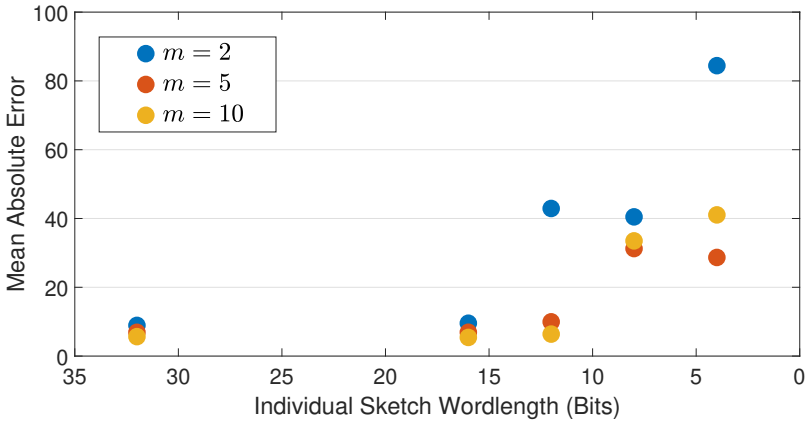

Figure 5.19: *Mean absolute error of the reconstructions for a sketch of size* $m = 2, 5$ *and* 10 *for individual sketch entries of 4,8,12,16 and 32 bit precision.*

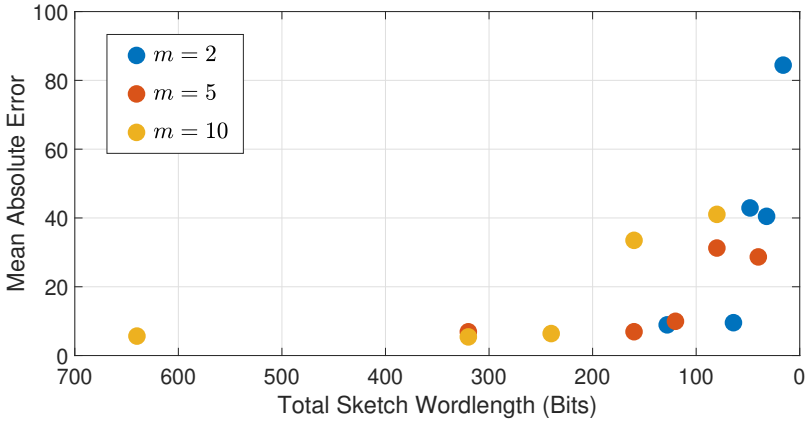

Figure 5.20: *Mean absolute error of the reconstructions for a sketch of size* $m = 2, 5$ *and* 10 *for the total wordlength of the whole sketch.*

5.5 Concluding Remarks

In this chapter, we proposed a novel sketching solution to handle the major data processing bottleneck of single-photon lidar caused by the fine resolution of modern high rate, high resolution ToF image sensors. Our approach involved sampling the characteristic function of the observation model to form online statistics that have dimensionality proportional to the number of parameters of the model. Furthermore, we developed an efficient sketching algorithm, inspired by ECF estimation techniques, which has space and time complexity that fundamentally scales with the size of the sketch m, and is independent of both photon count and depth resolution. Two sampling schemes are proposed that sample in regions of the characteristic function that are *blind* to photons originating from background sources. As a result, our method obtains estimates of the location and intensity parameters that are unbiased. Our novel sketch based acquisition removes the trade-off between depth resolution and data transfer complexity that is apparent in existing methods. Here we have only considered a simple pixel-wise depth estimate method in the form of the sketched MLE. However in the next chapter we demonstrate that it is straightforward to incorporate the sketched statistics into more sophisticated state-of-the-art multipixel reconstruction algorithms, such as the real-time 3D algorithm in [3] due to the Gaussian nature of the sketch statistics seen in Section 5.3.2.

5.A Deriving the Circular Mean Estimate From the ECF Estimation

Given a single frequency $\omega \in \mathbb{R}$, we can define the sketch as $z_N = \frac{1}{N}\sum_{j=1}^{N} e^{\mathrm{i}\omega x_j}$ and the goal is to solve:

$$\hat{\theta} = \arg\min_{\theta}(z_n - \Psi_{\mathcal{P}}(\omega))^2. \tag{5.22}$$

Clearly, Eqn 5.22 is minimised when $\Psi_{\mathcal{P}}(\omega) = z_n$ and equating the real and complex components we get:

$$\alpha e^{\frac{(\omega t)^2}{2}} \cos(\omega t) - (1-\alpha) D_{\frac{T-1}{2}}(\omega) = \frac{1}{N} \sum_{j=1}^{N} \cos(\omega x_j) \tag{5.23}$$

$$\alpha e^{\frac{(\omega t)^2}{2}} \sin(\omega t) = \frac{1}{N} \sum_{j=1}^{N} \sin(\omega x_j). \tag{5.24}$$

Notably, we can optimally choose the frequency to be $\omega = \frac{2\pi}{T}$ resulting in $D_{\frac{T-1}{2}}(\omega) = 0$ and thereby ensure the characteristic function is sampled in a region where the background noise is not present. Consequently, dividing (28) by (29) we get

$$\frac{\alpha e^{\frac{(\frac{2\pi t}{T})^2}{2}} \cos\left(\frac{2\pi t}{T}\right)}{\alpha e^{\frac{(\frac{2\pi t}{T})^2}{2}} \sin\left(\frac{2\pi t}{T}\right)} = \frac{\sum_{j=1}^{N} \cos(\omega x_j)}{\sum_{j=1}^{N} \sin(\omega x_j)}, \tag{5.25}$$

resulting in an optimal estimate of

$$\theta^* = \frac{T}{2\pi} \operatorname{phase} \left\{ \sum_{j=1}^{N} \cos\left(\frac{2\pi x_j}{T}\right) + \mathrm{i} \sum_{j=1}^{N} \sin\left(\frac{2\pi x_j}{T}\right) \right\} \tag{5.26}$$

5.B Photon Starved Regime

We evaluate the performance of our proposed sketched lidar method in the photon starved regime in comparison to transferring the photon time-stamps directly off-chip and estimating the depth of the surface. For fair comparison, we let $2m = N$ for each photon count N in the photon starved regime. Both the cross correlation and maximum peak estimate the depth location using the full photon count. Here we simulate a pixel of a lidar scene with a time window of $T = 100$ using a Gaussian IRF with pulse width $\sigma = 0.03T$ for photon counts $N = [1, 3, 5 \ldots, 15]$ and SBR varying between 0.01 and 100. For each photon count and SBR pair, 1000 Monte-Carlo simulations were executed with randomly chosen depth position $t_0 \in [1, 2, \ldots, T]$ and the RMSE was calculated.

Furthermore, we use the RMSE ratio between the sketched lidar and cross correlation depth estimation, defined as

$$R = \frac{\text{RMSE}_{\text{sketch}}}{\text{RMSE}_{\text{XCORR}}}, \tag{5.27}$$

where $\text{RMSE}_{\text{sketch}}$ and $\text{RMSE}_{\text{XCORR}}$ denote the RMSE of the sketched lidar and cross corre-

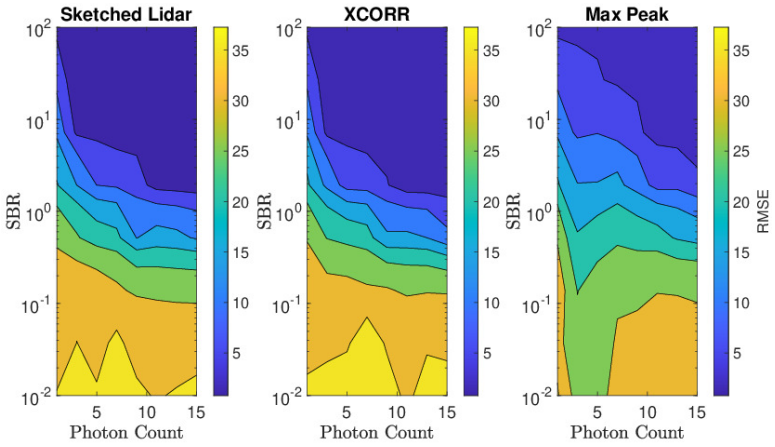

Figure 5.21: *Sketched Lidar performs comparatively well (in terms of RMSE) compared with the full data approaches of cross correlation (XCORR) and maximum peak estimation in the photon starved regime.*

lation estimation, respectively. An $R > 1$, indicates that the cross correlation achieves on average a smaller RMSE than sketched lidar. Similarly an $R < 1$, indicates the sketched lidar estimation achieves on average a smaller RMSE than cross correlation approach. Figures 5.21 and 5.22 show that the proposed sketched lidar approach does not suffer from a drop in estimation performance in both the photon starved regime and in the case of extremely low SBR in comparison with the cross correlation that estimates the depth using all the detected photons.

5.C Comparison to the iFFT approach

In Section 5.4.2, we compared our proposed sketched lidar approach to the iFFT approach. The iFFT approach cannot incorporate information about the impulse response function while in the sketched lidar method the impulse response function is integrated throughout. To demonstrate this, we compare the performance of the sketched lidar and iFFT techniques for the non-Gaussian asymmetric IRF used in Section 5.4.3.1 (See Figure 5.13). For a signal-to-background ratio varying between 0.1-100 and a photon count ranging between 10-1000, a pixel from a lidar scene was simulated with randomly chosen depth position between $1, \ldots, T$. A total of 1000 Monte-Carlo experiments were simulated for each SBR/photon count value with the RMSE recorded. For fair comparison we include an asymmetric correction for the iFFT approach to offset the bias of the asymmetric impulse response function. In Figure 5.23, the ratio between

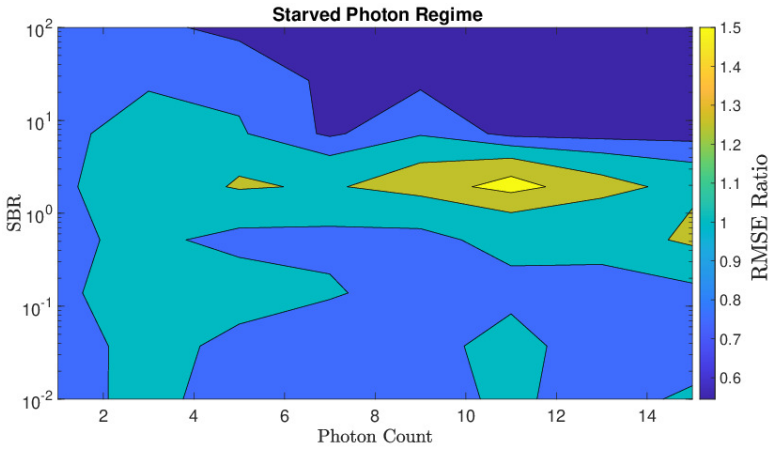

Figure 5.22: *Comparison of the depth reconstruction of sketched lidar and cross correlation (XCORR) using the RMSE ratio R for varying SBR levels and photon counts in the photon starved regime. Sketched Lidar performs favourably compared to XCORR for the majority of SBR values.*

the RMSE of the sketch results and the RMSE of the iFFT estimation, for e.g.:

$$R = \frac{\text{RMSE}_{\text{sketch}}}{\text{RMSE}_{\text{iFFT}}} \tag{5.28}$$

is displayed for $m = 2$.

The improvement using the sketched lidar method over the iFFT approach is apparent. For the majority of the SBR/photon count pairs the sketched lidar method achieves approximately half the RMSE of that of the iFFT approach, highlighting the lack of information of the IRF the iFFT approach has incorporated into its depth estimation.

s

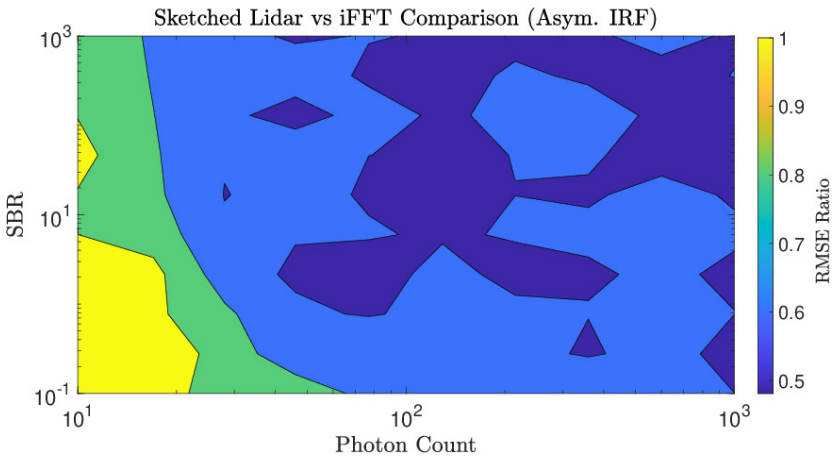

Figure 5.23: *Comparison of the depth reconstruction of sketched lidar and the iFFT method using the RMSE ratio R for varying SBR levels and photon counts for a non-Gaussian asymmetric IRF. Sketched Lidar performs equally or favourably to iFFT for all SBR and photon count pairs.*

Chapter 6
Robust Detection and Real-Time Processing of Sketched Single-Photon Counting Lidar

6.1 Introduction

In the last chapter, we developed a compressive learning framework for single-photon counting lidar that vastly reduced the volume of data needed to be transferred. The main aim was to set out the fundamentals of forming a sketch of the time-of-flight data that contained sufficient information to estimate the parameters of the lidar observation model. In this chapter, we extend the framework to include (i) a robust surface detection algorithm that detects the presence of a surface solely from the sketch, and (ii) a sketched real-time 3D algorithm which incorporates powerful point cloud denoisers that produces spatially regularized reconstructions.

A crucial aspect to a lidar pipeline is pixelwise detection of surfaces. Surface detection is often used to reduce both the data transfer and computational complexities of processing large point clouds. Regions of a scene may contain pixels that have no surface or object present. This is often the case in outdoor 3D imaging applications where the target may represent only a subset of the pixels. Removing the proportion of the pixels that contain zero surfaces prior to estimation can significantly reduce the complexities associated with data transfer and computation. However, current state-of-the-art surface detection algorithms have computational complexity that scales at best with $\mathcal{O}\left(T \log T\right)$ where T is the number of temporal bins in the TCSPC histogram. In this chapter, we propose a sketch-based surface detection algorithm that detects the presence of a surface or object solely from the sketch, leading to a reduced computational complexity of $\mathcal{O}(m)$.

Due to the spatial regularity of natural scenes, the parameters of the lidar observation model admit strong correlation in neighbouring pixels. This prior knowledge is exploited by several 3D reconstruction algorithms to improve the reconstruction quality over simple pixelwise depth

estimation methods. However, most of these state-of-the-art reconstruction algorithms have computational and memory complexity proportional to either the number of recorded photons or the depth resolution. This complexity hinders their real-time deployment on modern lidar arrays (see Section 2.5) which acquire potentially billions of photons per second. Using the framework set out in Chapter 5, we show that it is possible to modify these existing algorithms to require only the sketch. As a result, we can design high quality reconstruction algorithms that have computational and memory complexities that are proportional to the size of the sketch. Below we state the main contributions of this chapter.

- We propose a robust surface detection algorithm that forms a statistical hypothesis test directly on the computed sketch and exhibits a computational complexity of $\mathcal{O}(m)$. We again exploit the spatial correlation of natural scenes by incorporating a total variation (TV) regularizer that promotes a more homogeneous detection map of present targets and reduces the detection of spurious non-informative peaks.

- We propose a sketched real-time 3D (SRT3D) reconstruction algorithm that exploits the spatial correlation of natural scenes to improve the reconstruction quality compared to the original pixelwise sketched lidar algorithm introduced in Chapter 5.

- Using both real and synthetic datasets, we compare the point cloud estimation quality with other 3D reconstruction algorithms and demonstrate that our proposed SRT3D method is robust to challenging low SBR, low photon count scenes and achieves reconstructions that are competitive with the state-of-the-art.

This chapter is based of the surface detection work in [130] that appeared in the IEEE EUSIPCO conference 2021 and the multipixel sketched lidar work [131] that appeared at IEEE ICASSP conference 2022. The rest of the chapter is organized as follows. In Section 6.2, the sketch-based surface detection algorithm is proposed and we evaluate its robustness in challenging low SBR, low photon regimes using synthetic and real datasets. In Section 6.3 we introduce the SRT3D algorithm and analyse the reconstruction performance on both synthetic and real datasets. In Section 6.4, we finalize the chapter with some concluding remarks.

6.2 Pixelwise Surface Detection

This chapter begins with surface detection which is a crucial aspect of any lidar pipeline. If there are no objects present in the line-of-sight of the lidar device (e.g. outdoor setting), the recorded pixels will only consist of photon detections corresponding to background illumination which equates to $K = 0$ and $\alpha_0 = 1$ in Eqn 2.93. Figure 6.1 depicts a TCSPC histogram of a pixel that contains only background sources. Detecting and discarding pixels without peaks can avoid estimating non-existing surfaces, while reducing the computational load of posterior depth estimation. Figure 6.2 demonstrates a 3D image of a face with a corresponding detection map which shows if a surface is present or not in each pixel. In this example, 58% of pixels originate from background sources and can be discarded before posterior depth estimation reducing the overall complexities of processing.

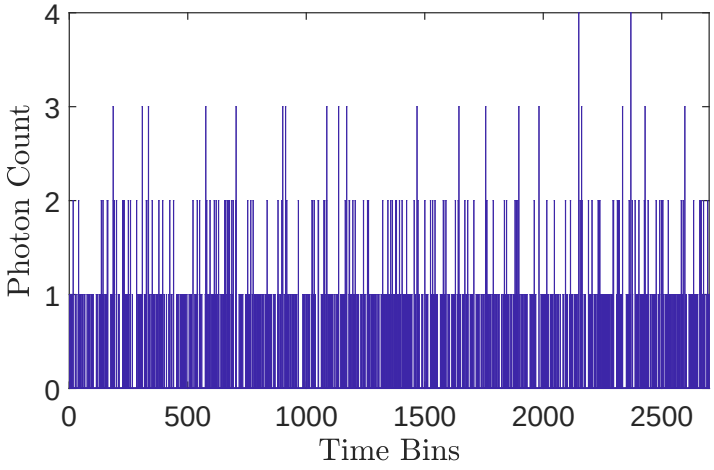

Figure 6.1: *A TCSPC histogram of a pixel containing no informative surface peak.*

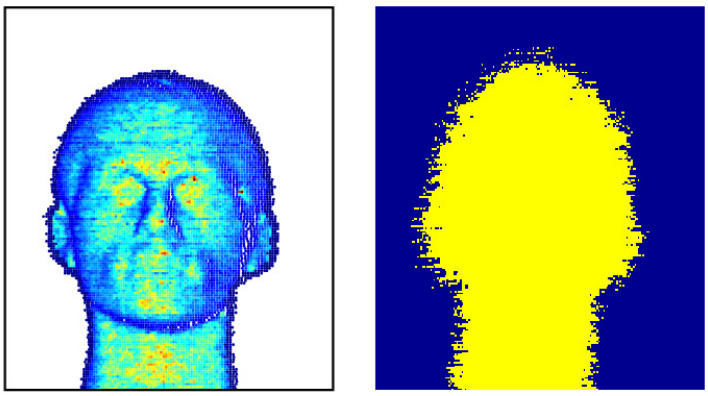

Figure 6.2: *A 3D image of a face (left) with its associated pixelwise surface detection map (right).*

Several existing surface detection algorithms [2, 132, 3] form a decision rule based on rejecting or accepting a hypothesis on a candidate observation model. By H_0, we define the null hypothesis by

$$H_0 : \mathcal{P}(x) = \mathcal{P}_b(x) \tag{6.1}$$

and the alternative hypothesis H_1 by

$$H_1 : \mathcal{P}(x) \neq \mathcal{P}_b(x) \tag{6.2}$$

In other words, the null hypothesis states that the observation model is equal to the distribution of background sources.

When a TCSPC histogram approach is used, the decision rule in Eqn 6.1 is equivalent to testing if the photons are distributed according to a homogeneous Poisson process. Under the inter-arrival time, the statistic $\Delta x = x_{i+1} - x_i$ is distributed according to an exponential random variable with parameter N/T, i.e.,

$$\Delta x \sim \mathcal{E}\left(\frac{N}{T}\right) \coloneqq \frac{N}{T} e^{-\frac{N}{T}\Delta x}, \tag{6.3}$$

where $\mathcal{E}$ denotes the exponential distribution. Hence, a standard test consists of computing the Kolmogorov-Smirnov (K-S) statistic using the empirical inter-arrival time distribution [133]. However, this test has important drawbacks. First, the statistic requires storing all the time-stamps, scaling linearly in the number of collected photons N or histogram size T. Secondly, the test cannot account for the discrete nature of the time-stamps collected by the TCSPC device.

An alternative method amenable to discrete time-stamps consists in checking whether the photon count in all $T_r \leq T$ bins of a coarse histogram have a mean close to N/T_r (the expected number of photons under H_0), using a χ-squared test. In this setting, if T_r is small, small peaks can be hidden in the coarse depth resolution, hindering the detection and posterior depth estimation performance. On the other hand, if T_r is too large, a small number of photons per bin would depart significantly from the Gaussianity assumption of the χ-squared test, degrading the performance of the method. This trade-off is shown in the experiments in Section 6.2.3. State-of-the-art techniques in lidar detection, which we compare to in Section 6.2.3, use a Bayesian approach by asserting priors onto the background levels and employ a decision rule similar to Eqn 6.1. However, the computational complexity scales at best with $\mathcal{O}\left(T \log T\right)$ and the whole

histogram is required in memory.

6.2.1 Sketch-Based Detection Algorithm

Here we propose a sketch-based detection scheme based solely on the compact sketch. First note that due to the one to one correspondence between probability distributions and their corresponding characteristic functions (see Section 2.1.3.1), the hypothesis test in Eqn 6.1 can be equivalently defined by the null hypothesis H_0'

$$H_0' : \Psi_{\mathcal{P}}(\omega) = \Psi_{\mathcal{P}_b}(\omega) \tag{6.4}$$

and the alternative hypothesis H_1'

$$H_1' : \Psi_{\mathcal{P}}(\omega) \neq \Psi_{\mathcal{P}_b}(\omega). \tag{6.5}$$

Next, recall from Section 5.3 that a sketch converges to a Gaussian distribution

$$\mathbf{y}_N \xrightarrow{\text{dist}} \mathcal{N}\left(\mathbf{y}_\theta, N^{-1}\Sigma_\theta\right) \tag{6.6}$$

where for convenience we denote $\mathbf{y}_\theta = [\Psi_{\mathcal{P}}(\omega_j)]_{j=1}^m$ and we recall $\Sigma_\theta \in \mathbb{C}^{m\times m}$ has entries $(\Sigma_\theta)_{ij} = \Psi_{\mathcal{P}}(\omega_i - \omega_j) - \Psi_{\mathcal{P}}(\omega_i)\Psi_{\mathcal{P}}(-\omega_j)$ for $i, j = 1, 2, \ldots, m$. Denote by D^2 the test statistic defined by

$$D^2 \coloneqq N(\mathbf{y}_N - \mathbf{y}_\theta)^T \Sigma_\theta^{-1}(\mathbf{y}_N - \mathbf{y}_\theta), \tag{6.7}$$

then it can be seen [134, 135] that the test statistic D^2 follows a χ-squared distribution:

$$D^2 \xrightarrow{\text{dist}} \chi_\nu^2 \tag{6.8}$$

with $\nu = m-p$ degrees of freedom where p is the number of parameters of the lidar observation model $\mathcal{P}$. Under the null hypothesis H_0', it can easily be seen from Eqn 6.6 that $\mathbf{y}_\theta = \mathbf{0}_m$ and $\Sigma_\theta = \mathbf{I}_m$ where $\mathbf{0}_m$ and $\mathbf{I}_m$ denote the m dimensional zero vector and the $m \times m$ identity matrix, respectively. Hence, the test statistic D^2 simply reduces to

$$D^2 = \|\mathbf{y}_N\|_2^2. \tag{6.9}$$

Under the null hypothesis, the lidar observation model ($\mathcal{P} = \mathcal{P}_b(x_p) = 1/T$) has $p = 0$ parameters, therefore D^2 follows a χ-squared distribution with $\nu = m$ degrees of freedom. One can therefore reject the null hypothesis H_0' at significance level β if $D^2 > \bar{z}_\beta$ where $\bar{z}_\beta$ is the upper β-percentile of the χ^2_m distribution [135]. A summary of the sketch-based surface detection scheme is detailed in Algorithm 5. Importantly, the decision rule is based solely on the sketch of size m. This is significant as (i) the full data of the TCSPC histogram is not required in the computation and can be discarded from memory (ii) the squared test statistic can be computed in $\mathcal{O}(m)$.

Algorithm 5 Pixelwise Sketch-based Surface Detection Algorithm

Require: Sketch $\mathbf{y}_N$, significance level β.
  Compute test statistic $D^2 = \|\mathbf{y}_N\|_2^2$.
  Compute upper β- percentile of χ^2_m distribution $\bar{z}_\beta$.
  if $D^2 > \bar{z}_\beta$ **then**
    Reject H_0' at significance level β and detect presence of a surface.
  else
    Accept H_0'.
  end if

In some practical settings, the distribution of background photons $\mathcal{P}_b$ might not be exactly constant. This is often attributed to the so-called pile-up phenomenon whereby the dead-time of the SPAD is too slow to process successive photon detection events, and therefore the background noise becomes non-constant [136]. In these cases, the test statistic D^2 can be easily modified to account for a data-driven $\hat{\mathcal{P}}_b$, using background photons collected in a calibration step,

$$\hat{\mathbf{y}}_N = \mathbb{E}_{\hat{\mathcal{P}}_b}\{\Phi(x_i)\}. \tag{6.10}$$

The test statistic is then $D^2 \coloneqq \|\mathbf{y}_N - \hat{\mathbf{y}}_N\|_2^2$. It is worth noting that the data-driven test can also be interpreted as a random features version of the maximum mean discrepancy [137] as discussed in Section 2.3.2.3.

6.2.2 Spatial Regularization

Neighbouring pixels in a lidar scene typically exhibit the same number of surfaces owing to spatial correlation. Exploiting the inherent spatial correlation in typical lidar scenes can further reduce the occurrence of false positives. In [132], Tachella et al. proposed a total variation (TV) based spatial regularization that created a more homogeneous map of the present targets.

Here we include a similar spatial regularization based on the goodness-of-fit. Formally, the TV based spatial regularization is defined by the map

$$\hat{\boldsymbol{v}} \coloneqq \mathcal{H}_{0/1}\left(\arg\min_{\boldsymbol{v}} \|\boldsymbol{v}-\boldsymbol{u}\|_2^2 + \tau\|\boldsymbol{v}\|_{\text{TV}}\right) \tag{6.11}$$

where the input image $\boldsymbol{u}$ contains the χ-squared statistic D^2 of pixel (i,j), $\|\cdot\|_{\text{TV}}$ is the isotropic TV operator, τ is a user-defined regularization parameter and $\mathcal{H}_{1/0}$ is a hard-thresholding operator which assigns 1 to positive inputs and 0 otherwise. In Section 6.2.3, it is demonstrated that the added spatial regularization can help remove a proportion of false positive alarms producing a more homogeneous detection map.

6.2.3 Empirical Results

In this section, we evaluate the sketch-based detection scheme on both real and synthetic data. First, we analyse the effect of the signal-to-background ratio (SBR), defined by SBR$= \alpha/(1-\alpha)$, and the photon count N on both the true positive and false alarm rate, using a Gaussian impulse response with standard deviation $\sigma = T/100$, for $T = 5000$. Figure 6.4 shows a map of the empirical probability of detecting a single peak for various SBRs and photon counts for the proposed sketch-based detection. Even for moderately high SBR, for example SBR$= 1$, the detection scheme only requires approximately 20 photons to achieve high probability of detecting a single peak. Figure 6.3 shows the SBR/photon count level-curves for a true positive rate of 95% for various sketch sizes and full-data approaches. For the full-data approach, a χ^2 test was constructed on the true observation model in Eqn 2.93 where adjacent bins were concatenated to maximise the power of the hypothesis test. For reference, we also include the K-S test discussed in Section 6.2 which is equivalently performed on the full data (see [133] for details). For each test the significance level was set at $\beta = 0.05$. Figure 6.5 depicts the empirical probability of false alarm (PFA) as a function of the photon count for various sketch sizes and for the aforementioned full data hypothesis tests.

Next, we compare the proposed sketch-based detection algorithm with the χ^2 test on the full data observation model as well as the two detection methods proposed by Tachella et al. in [132], using a real lidar dataset consisting of a polystyrene head measured at a stand-off distance of 325 metres. The dataset is different from the polystyrene dataset from the empirical results of Section 5.4 and Section 6.3.1 as the scene was captured outdoors resulting in pixels containing

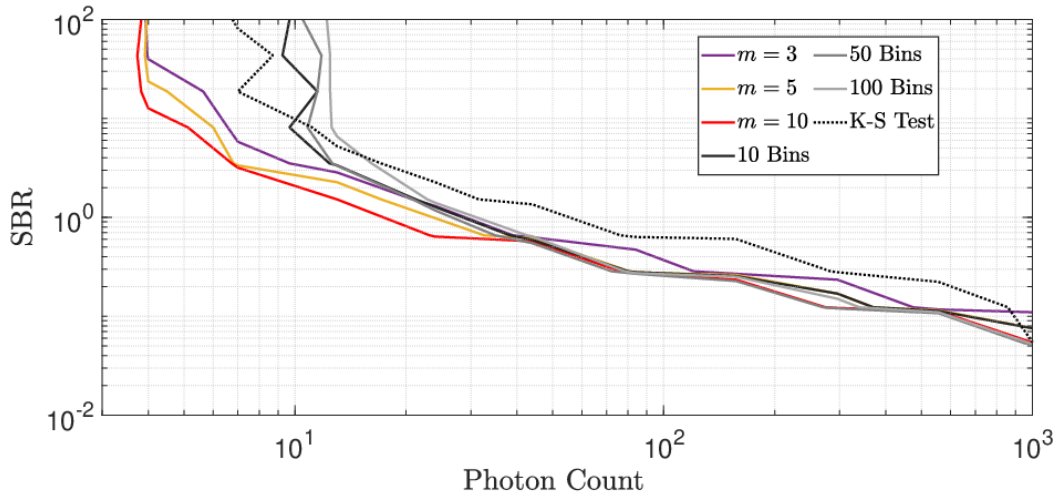

Figure 6.3: *Detection performance of the sketch-based method for sketch sizes of* $m = 3, 5, 10$*, the coarse histogram test for histograms of size* $T_r = 10, 50, 100$*, and the full data K-S test. The graphs correspond to a detection probability of 95%.*

no targets as well as exhibiting a smaller SBR. See details of the dataset in [2]. The dataset consists of 200×200 pixels with $T = 2700$ histogram bins per pixel and an approximate SBR of 0.29. Figure 6.6 shows the detection maps for two different per-pixel acquisition times (30 ms and 3 ms) corresponding to an average photon count of 900 and 90 photons, respectively. The sketch size was set at $m = 5$ and the significance level was set at 0.05 and 0.2 for the 30 ms and 3 ms acquisition times, respectively. Also included is the proposed sketch-based method with spatial TV regularization as discussed in Section 6.2.2. The PD and the PFA for each acquisition time are shown in Table 6.1 for each detection scheme. The PD and PFA for both the sketch and sketch plus TV regularization are depicted in Figure 6.7 for increasing sketch size m.

The results show that on both synthetic and real datasets the sketch-based detection scheme achieves a similar, or better, PD/PFA trade-off than the full data χ^2 detection test. In fact the sketch-based detection scheme achieves a far lower PFA than the full data χ^2 detection for both acquisition times. In comparison to the state-of-the-art results by Tachella et al [132], the sketch and sketch plus TV regularization produce a depth map that is competitive. For instance, for the longer acquisition window of 30ms, the sketch plus TV achieves a PD and PFA trade-off of (96.6%, 0.9%) compared with the TV regularized version of [132] that achieves a similar tradeoff of (98.4%, 3.5%). Notably, the methods of Tachella et al. exhibit a computational complexity of $\mathcal{O}(T \log T)$ compared to our sketch-based approach that scales with $\mathcal{O}(m)$ where,

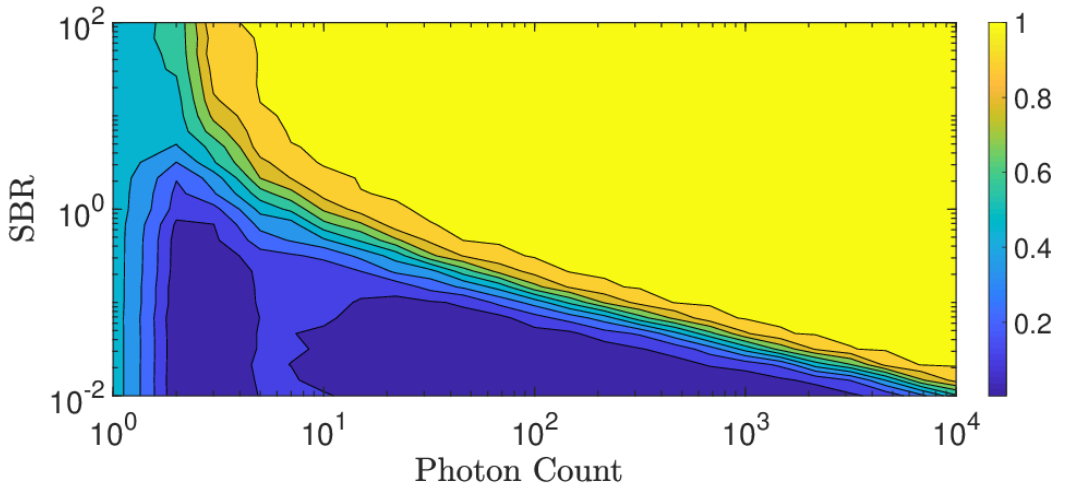

Figure 6.4: *Empirical probability of detection for the proposed sketch-based detection scheme using a sketch size $m = 10$.*

	PD %		PFA %	
	30ms	3ms	30ms	3ms
Tachella et al.	98.5	82.3	0.7	6.5
Tachella et al. (TV)	98.4	93.7	3.5	1.1
K-S Test (Full Data)	100	49.9	99.3	6.9
Hist. (50 Bins)	97.5	77.6	8.8	19.9
Sketch	95.4	77.2	1.4	14.4
Sketch + TV	96.6	88.1	0.9	0.5

Table 6.1: *Probabilities of detection (PD) and probabilities of false alarm (PFA) for the proposed sketch-based detection schemes and other detection algorithms. The sketch size is set at $m = 5$ and the full data χ^2 test was chosen using 50 adjacent bins to optimise the PD/PFA trade-off.*

in this example, $T = 2700$ and $m = 5$ respectively. Furthermore, as discussed in Section 6.2, the full data K-S test struggles to account for the discrete nature of the time-stamps and detects a surface for nearly all pixels in the scene for the longer 30ms acquisition time.

6.2.4 Discussion

Throughout the whole of Section 6.2, we focus on the hypothesis test in Eqn 6.1 that is based on if there is or isn't a surface present for a given pixel as this is of particular importance for downstream lidar tasks. However, recall that the circular mean solution in Eqn 5.3 provides a closed form expression for the parameters of single peak observation model (e.g. $K = 1$). In

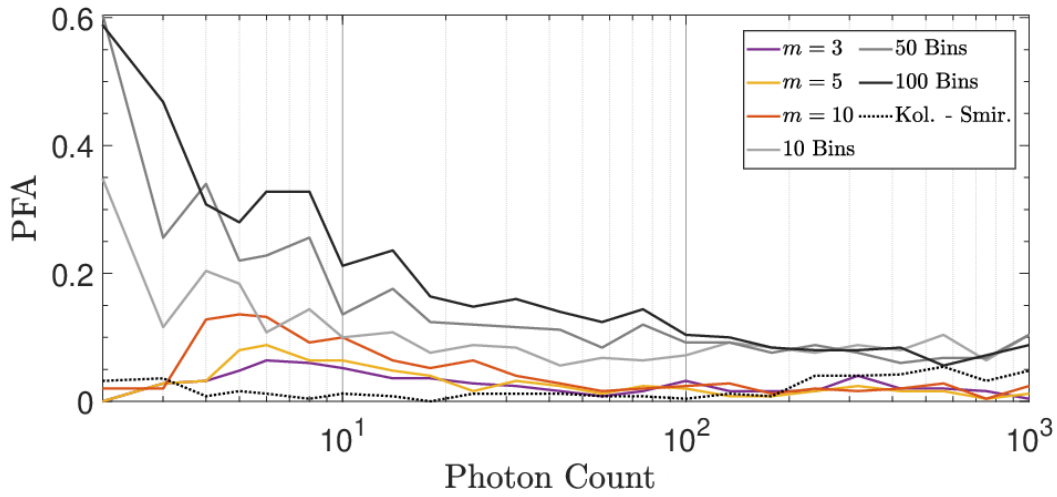

Figure 6.5: *Probability of false alarm of the sketch-based method for sketch sizes of* $m = 3, 5, 10$*, the coarse histogram test for histograms of size* $T_r = 10, 50, 100$*, and the full data K-S test.*

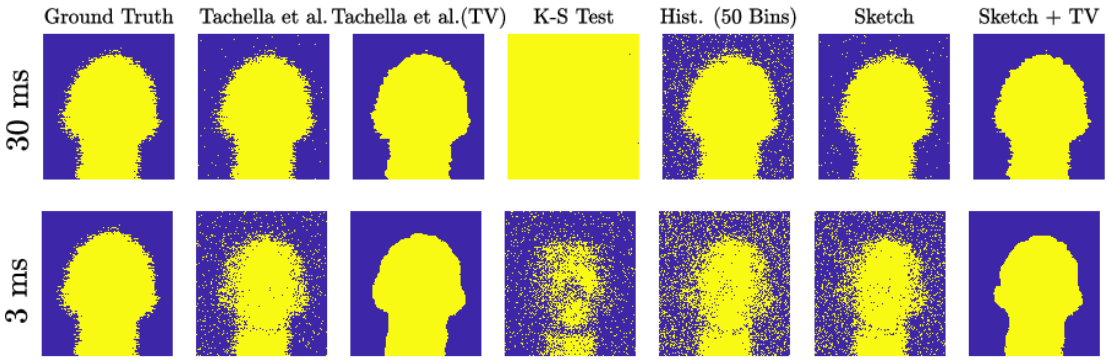

Figure 6.6: *Detection maps of the polystyrene dataset [2] for the proposed sketch and sketch plus TV detection schemes in comparison with other non-compression detection techniques.*

many lidar scenes, pixels containing only $K = 1$ surfaces form a substantial subset of the whole field of view. As the circular mean can be computed in closed form, it has a computational complexity of $\mathcal{O}(1)$. By detecting pixels containing $K = 1$ surfaces, one can instantly estimate the associated parameters using the circular mean solution without having to resort to using one of the sketch estimation algorithms we have proposed so far in this thesis, thereby reducing the computational complexity of processing the point cloud further. Subsequently, the second hypothesis test can be defined by

$$\tilde{H}_0 : \Psi_{\mathcal{P}}(\omega) = \alpha \Psi_{\mathcal{P}_s}(\omega) \; + \; (1 - \alpha) \Psi_{\mathcal{P}_b}(\omega) \tag{6.12}$$

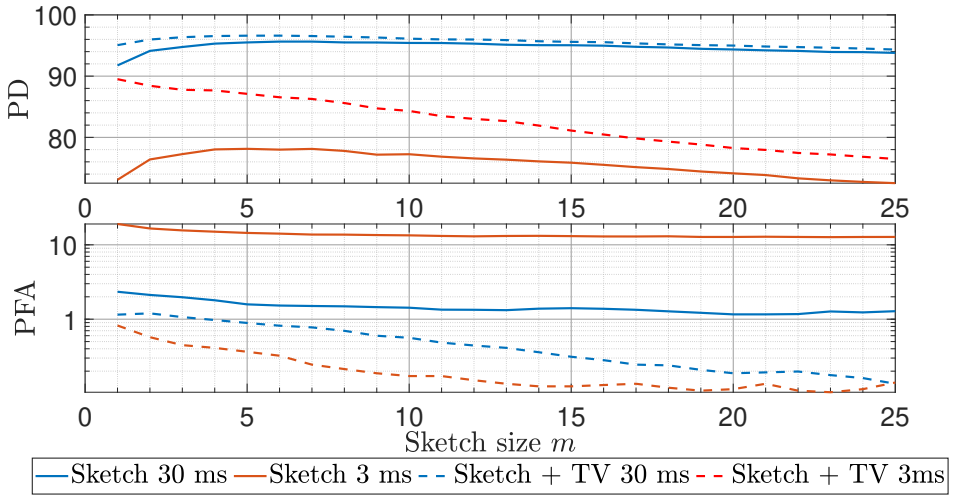

Figure 6.7: *Empirical probabilities of detection (top) and false alarm (bottom) for the evaluated detection methods using the polystyrene head dataset.*

and the alternative hypothesis $\tilde{H}_1$ by

$$\tilde{H}_1 : \Psi_{\mathcal{P}}(\omega) \neq \alpha \Psi_{\mathcal{P}_s}(\omega) \; + \; (1-\alpha)\Psi_{\mathcal{P}_b}(\omega) \tag{6.13}$$

In this instance, the test statistic D^2 can computed as in Eqn 6.7 and follows a χ-squared distribution with $\nu = m - 2$ degrees of freedom due to the $p = 2$ parameters (i.e. α, t) of the null hypothesis observation model. Moreover, it is shown in [134, 135] that one only needs a consistent estimator (see Definition 1) of the true lidar observation parameters θ to form the test, for instance the circular mean estimate which can be computed in $\mathcal{O}(1)$. As a result, as long as one has a consistent estimate of θ, the hypothesis test is not sensitive to how it is initialised. We leave implementation of this extended hypothesis test to future work.

6.3 Multipixel Sketched Lidar

In Chapter 5, we proposed a sketched lidar framework for single-photon counting lidar which focused on per pixel depth estimation. Due to the spatial regularity of natural scenes, parameters in neighbouring pixels are generally strongly correlated. This prior knowledge is exploited by several 3D reconstruction algorithms [138, 63, 139, 140, 123, 141, 127] to improve the quality with respect to simple pixelwise depth estimation. Recall from Section 5.4.1.1 that single-photon lidar devices acquire an array of $N_r \times N_c$ pixels. By encompassing all the parameters in a scene into $\boldsymbol{\theta} = (\theta_{1,1}, \ldots, \theta_{N_r,N_c})$, then most algorithms solve the following optimization

problem [61]

$$\arg\min_{\boldsymbol{\theta}} \sum_{i,j}^{N_r,N_c} f_{\boldsymbol{z}_{i,j}}(\theta_{i,j}) + \mathfrak{R}(\boldsymbol{\theta}) \tag{6.14}$$

where $\boldsymbol{z}_{i,j}$ denotes the observed histogram at the (i,j)th pixel, $f_{\boldsymbol{z}_{i,j}}(\theta_{i,j})$ are per-pixel data fidelity terms (negative log-likelihood of the ToF list or histogram observation models (see Section 2.5)), and $\mathfrak{R}(\boldsymbol{\theta})$ is a spatial regularization term which encodes the prior information about the spatial regularity of typical scenes. There has been significant efforts dedicated to the design of powerful regularizations $\mathfrak{R}(\boldsymbol{\theta})$. The RT3D algorithm [3] exploits the plug-and-play framework [142] together with a fast computer-graphics point cloud denoiser to design a regularizer that can capture the geometry of complex scenes while also simultaneously selecting $K \geq 1$ surfaces per pixel. However, existing algorithms (including RT3D) require multiple evaluations of the data fidelity terms, and thus suffer from large memory requirements and a computational complexity which is at least linear in the number of photon detections or histogram bins [3].

In this chapter, we propose to replace the histogram-based loss in Eqn 6.14 for the more compact sketch cost function in Eqn 5.10, while leveraging the spatial regularization penalty of existing methods. The proposed objective can be expressed as

$$\arg\min_{\boldsymbol{\theta}} \sum_{i,j}^{N_r,N_c} N_{i,j} \|\mathbf{y}_{i,j} - \Psi_{\theta_{i,j}}\|^2_{\mathbf{W}_{i,j}} + \mathfrak{R}(\boldsymbol{\theta}) \tag{6.15}$$

where $\mathbf{y}_{i,j}$ is the sketch associated with the (i,j)th pixel. The number of detected photons $N_{i,j}$ controls the trade-off between the data-fidelity and regularization terms. As the number of detected photons increases, the data fidelity term dominates Eqn 6.15, which tends to the non-regularized problem in Eqn 5.10. In order to perform real-time reconstruction with an arbitrary number of photon detections, we propose a sketched version of the RT3D algorithm, which we name SRT3D. The proposed algorithm replaces the histogram-based likelihood of RT3D for the sketched loss of Eqn 6.15.

6.3.1 Experiments

We evaluate the proposed SRT3D algorithm on the two real datasets considered in Section 5.4: a polystyrene head at a distance of 40 metres [63] and a scene with two people walking behind a camouflage net at a distance of 320 metres [3]. We compare the proposed method with 3 other

algorithms: traditional cross-correlation [143] (XCORR), the pixelwise SMLE reconstruction proposed in Chapter 5 which doesn't exploit spatial regularization, and RT3D which accesses the full fine-resolution ToF data. For reduced computational load, we set the weighting matrix in Eqn 6.15 to the identity, i.e. $\mathbf{W}_{i,j} = \mathbf{I}_m$ for all i, j. Although this is less efficient statistically (see Chapter 5), the gains from the spatial regularizer outweigh this loss. All the experiments were performed using an NVIDIA RTX 3070 laptop GPU.

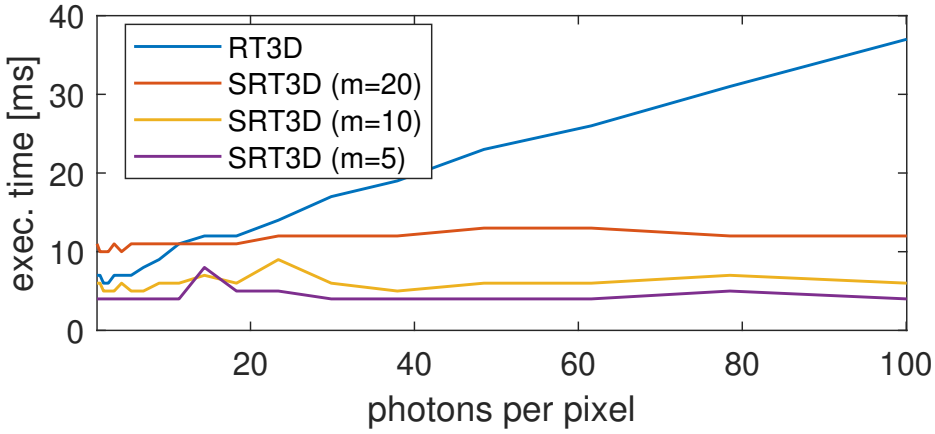

Figure 6.8: *Execution time of RT3D [3] and the proposed sketched SRT3D as a function of the mean number of photons per pixel. RT3D suffers from a linear complexity, whereas SRT3D only depends on the size of the sketch m.*

Recall from Section 5.4 that the polystyrene head dataset has size of 141×141 pixels with $T = 4613$. Most of the pixels in this scene contain exactly $K = 1$ surface. A ground-truth reference was obtained using using the standard cross-correlation algorithm on the raw ToF information (with high number of photons per pixel and high SBR). Using this reference and the observation model in Eqn 2.92, we synthesized multiple datasets for different mean photons per pixel n and SBR levels. Figure 6.9 shows the 3D reconstructions obtained for SBR levels of 10, 1 and 0.1. All methods perform similarly when the number of photons and SBR are large. Notably, a sketch of size $m = 5$ is sufficient to provide good reconstructions. However, when the scene contains a low number of photons or low SBR, pixelwise methods fail to provide good reconstructions, whereas both RT3D and SRT3D provide good reconstructions. In this challenging setting, a sketch of size $m = 10$ sufficiently provides a reconstruction that has the same quality as the ones obtained in the full data case. True and false detections, depth absolute error (DAE), and normalised intensity absolute error (IAE) (as defined in [3]) are presented in Figure 6.10 for an SBR of 1 and different number of photons per pixel.

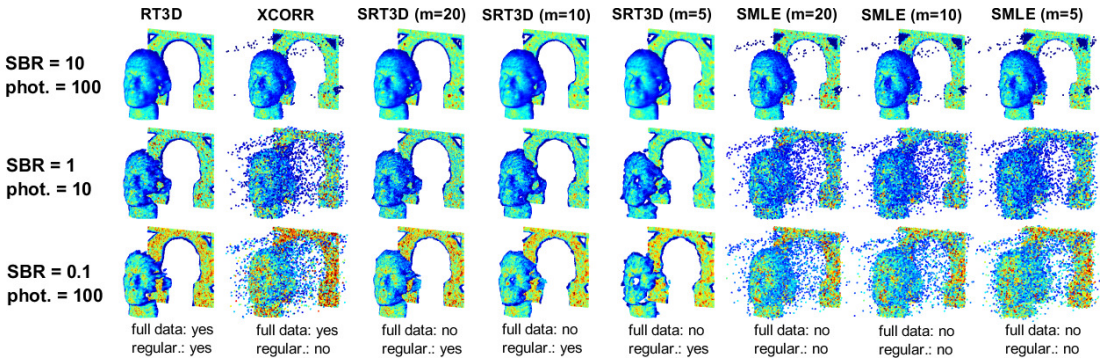

Figure 6.9: *3D reconstructions obtained by the proposed sketched RT3D algorithm for different sketch sizes m and other competing algorithms. The proposed SRT3D method incorporates spatial regularization, providing stable reconstructions in settings with low SBR or low number of measured photons.*

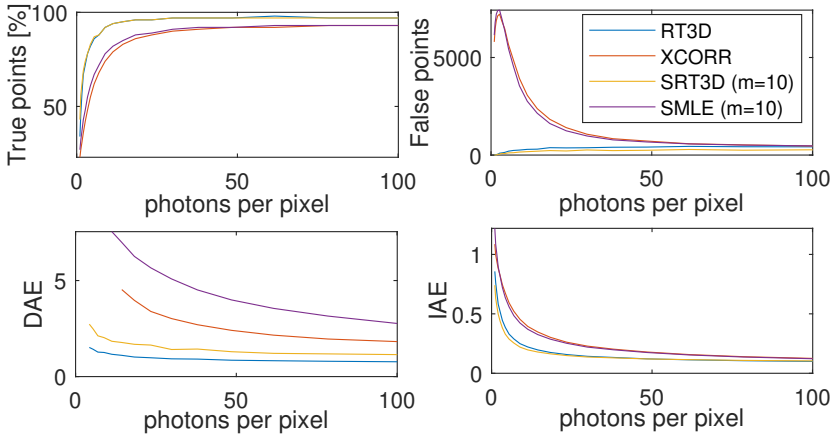

Figure 6.10: *Performance of the evaluated algorithms for the polystyrene head dataset with SBR=1.*

Figure 6.8 shows the execution time of RT3D and SRT3D as a function of the mean number of photons per pixel in the polystyrene head datasets. The GPU memory requirements of RT3D become prohibitive if the number of observed photons is in the order of hundreds per pixel per frame, whereas the sketched version has a complexity which is independent of the number of photons, and can handle any number of photons in real-time. Table 6.2 shows the execution time of SRT3D for increasing array sizes, demonstrating that the proposed method can process up to 705×705 arrays at 14 frames per second on a laptop computer. The datasets were generated by upsampling the head reference before the synthesis of photon detections.

Figure 6.11 shows the reconstructions of SRT3D and RT3D for single frame of the camouflage dataset in [3], which is composed of 32×32 pixels with $T = 153$. Recall from Section 5.4

m/pixels	141^2	282^2	423^2	564^2	705^2
$m = 5$	6	12	28	55	68
$m = 10$	7	18	35	60	88

Table 6.2: *Execution time in milliseconds for different scene sizes in pixels obtained by the proposed sketched RT3D algorithm for a sketch size of $m = 5$ and $m = 10$, respectively.*

that most of the pixels in the scene contain $K = 2$ surfaces, which makes the reconstruction task more challenging. However, a sketch of size $m = 10$ is sufficient to provide the same reconstruction quality as using the full 153 bins. Although the original fine resolution T is not large, the execution time of SRT3D for $m = 10$ was 12 ms, whereas for RT3D it was 20 ms.

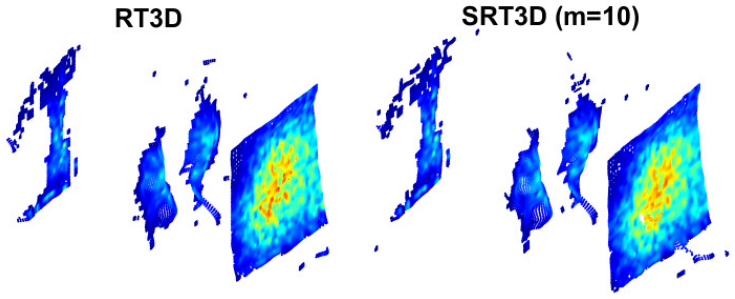

Figure 6.11: *Reconstruction of the scene in [3] with 2 surfaces per pixel by the original RT3D and its sketched version. Using a sketch size of only $m = 10$ is enough to provide the same reconstruction quality.*

6.4 Concluding Remarks

In this chapter, we developed the sketched lidar framework introduced in Chapter 5 to include both a robust surface detection algorithm based solely on the sketch and proposed a real-time sketched implementation that incorporated powerful regularizers $\mathfrak{R}(\boldsymbol{\theta})$. While our results focused on a sketched version of the RT3D algorithm [3], the ideas presented here can be used to develop sketched versions of other existing regularized lidar methods by simply replacing the data fidelity term $f_{\mathbf{z}_{i,j}}$ in Eqn 6.14 by the sketch cost function term in Eqn 5.10. Furthermore, we developed a detection scheme based solely on a compact representation sketch that is robust in detecting the presence of a surface for each pixel in the lidar scene. As a result, pixels consisting of non-existing surfaces can be discarded from memory reducing the overall computational and memory load of transferring and reconstructing a lidar scene. Moreover, it is shown that only a minimal sized sketched is required to achieve a high probability of detection

on both synthetic and real datasets, achieving a better PD/PFA trade-off than the corresponding χ-squared test on the original histogram data.

Chapter 7
Conclusion and Future Perspectives

7.1 Conclusions

In this work, the ultimate goal was to develop the compressive learning framework in [1, 6] by introducing new models and applications. From our extensive theory and algorithmic designs in Part I, to the development of a robust sketching framework to reduce the data transfer bottleneck of modern lidar device in Part II, we believe this has been demonstrated.

In Chapter 3, we developed a compressive ICA scheme that consisted of both theory and practical algorithms. At the core of the scheme is the existence of a low dimensional model set induced by the solutions to the cumulant based ICA problem. The solutions of the model set admitted a sparse tensor decomposition that we exploited to form a dimension reducing sketch and to develop both an iterative projected gradient and alternating steepest descent version of the compressive ICA algorithm. It was shown through theory and consolidated by phase transition experiments that a sketch of order $\mathcal{O}(n^2)$ is of sufficient size to retain enough salient information to accurately estimate the parameters of the ICA model. The compressive ICA scheme leads to substantial memory as existing cumulant based ICA methods require in memory either the full data matrix or the 4th order cumulant tensor of size $\mathcal{O}(Nd)$ and $\mathcal{O}(n^4)$, respectively.

The ICA model belongs to the larger class of semi-parametric models where their corresponding parameters are typically identified through a set of statistics associated with the data. In Chapter 4, we reformulated the existing CL framework to cater specifically for the inherent structure and topology admitted by semi-parametric models. The reformulation enables a clear blueprint in designing future compressive learning schemes for semi-parametric models and establishes early on if, or when, compression can be attained with respect to the model's underlying dimensionality. Through the use of a case study, we highlighted that a compressive GPCA scheme cannot always enable efficient compression in comparison to using the full data as the identifiable statistic (i.e. the correlation matrix of the Veronese embeddings) scales exponentially with the model set dimensions. Two key conclusions were drawn from this chapter: (1)

in its current form, by leveraging identifiable statistics, we can reduce the compressive learning problem down to a finite dimensional compressive sensing setting that is often easier to solve, and (2) the reliance on an identifiable statistic can be the poison as well as the medicine as it may not always exist for a given semi-parametric model, and if it does it may have a size that does not scale favourably with the model's underlying dimensions.

In Part II of the thesis, a sketching framework for reduced data transfer in modern day lidar devices was developed that circumvented the compression-resolution trade-off that is inherent in existing methods. The key conclusion of Chapter 5 was that the size of the sketch needed only to scale with the number of surfaces in the field of view and was fundamentally independent of both the temporal resolution and the photon count. As most scenes consist of $K = 0, 1$ or 2 surfaces, enormous compression was achieved with only negligible loss of information as demonstrated on both synthetic and real datasets. In contrast to the sampling schemes proposed in [6], we designed a deterministic sampling strategy that sampled the first m frequencies that were *blind* to the background noise.

In Chapter 6, the sketched lidar framework was extended by designing a plug and play multi pixel denoiser algorithm that was robust to both low SBR and small photon counts. Importantly, by replacing the data fidelity term in Eqn 6.14 with the sketch cost function term in Eqn 5.10 in state-of-the-art multi pixel denoisers, one can easily develop sketched versions of other existing and future regularized methods. Furthermore, a surface detection algorithm was proposed in Chapter 6 that was based solely on the sketch and admitted a computational complexity of $\mathcal{O}(m)$ compared with existing schemes that had a complexity of $\mathcal{O}(T \log T)$, enabling substantial computational and memory complexity compression. Several key conclusions were drawn from Chapter 6. The first being that sketches not only provide sufficient information for inference but they also allow for robust detection even at small sketch sizes. Secondly, the flexibility of the sketched lidar framework was demonstrated by substituting the compressive lidar cost function for the data fidelity term in the full data lidar denoisers. Incorporating well-established denoisers or regularizers of a learning task provides a possible avenue for future compressive learning implementation where the sparsity of the model set may not be fully understood.

7.2 Future Perspectives

The work presented in this thesis aimed to address a small amount of the gaps in compressive learning literature and intended on developing both new models and applications. Many interesting and challenging problems arose throughout the duration of the PhD studies. In this final section, some key open problems and possible future directions will be detailed.

7.2.1 Efficient One-Stage Compressive ICA

In Section 3.4.2, a one-stage compressive ICA approach was proposed that enabled parameter estimation without an initial prewhitening stage by modifying the existing 2 stage compressive ICA scheme. This is of particular interest in streaming applications where we might be limited to one instance of the data making a prewhitening stage impossible. However, a major limitation of our one-stage compressive ICA algorithm is requiring the calculation of $\mathbf{A}\bar{\mathbf{V}}^{-1}\operatorname{vec}(\mathscr{X})$ at each iteration of the algorithm. Recall that $\bar{\mathbf{V}} \coloneqq \mathbf{V} \otimes \mathbf{V} \otimes \mathbf{V} \otimes \mathbf{V} \in \mathbb{R}^{n^4 \times d^4}$, where $\mathbf{V}$ is the whitening matrix which can be estimated through the second order moments of the unwhitened sketch. For $d \approx n$, the extra computation does not substantially increase the complexity of the whole compressive ICA algorithm. However, in some cases the number of mixed components is much larger than the number of independent components, i.e. $d \gg n$, and the prewhitening stage acts as a form of dimensionality reduction in its own right. In this case, the polynomial dependency on d makes our current one-stage particularly slow for large d. In future work, we wish to develop a one-stage compressive ICA algorithm that is cheaper and efficient to run. This would require further exploration of the structure induced by the solutions of unwhitened 4th order moments so that we can design specific projection operators.

7.2.2 Alternative Compressive Semi-Parametric Learning

In Chapter 4, we reformulated the existing compressive framework so that it generalizes to semi-parametric models. However, a key limitation, as discussed earlier in this chapter, is the reliance of an identifiable statistic which may not be either readily available or scale favourably with the underlying dimensions of the model. An important line of research would be to explore other ways of forming a semi-parametric sketch that does not require some identifiable statistic. For instance, one could possibly sketch the equivalence class of distributions and design suitable decoders that enable accurate parameter estimation.

7.2.3 Compressive Transfer Learning

One of the main limitations of the current compressive learning approach is that the design of a sketch is heavily dependent on the task or model of interest where the construction of the sketching operator is highly non-trivial. In many cases, a given dataset is used to query multiple tasks and build several models. This would require a sketch to be designed and computed for each individual task. A future direction of research would be to design multitask sketches that can be computed once and be used for several tasks.

7.2.4 On-Chip Implementation of Sketched Lidar

In Chapter 5, we extensively tested our sketched lidar framework on several real datasets. However, the SMLE algorithm was executed on prerecorded photon counting data where traditional TCSPC histograms in Section 5.1 were converted to raw photon time-stamps. A large body of research needs to be undertaken to bridge the gap between simulations and practical implementation of the sketching approach within an FPGA on the lidar device. A significant challenge is the computation of the sinusoidal functions within the sketch. It is known that such functions are often expensive to compute from a logic resource point of view. In Section 5.2.3, we briefly discussed various efficient logic-based schemes for constructing the sketch on-chip. Schellekens et al. showed that in principle one can replace the sinusoidal functions of the sketch by alternative periodic functions (e.g. square waves or triangle waves) in conjunction with random dithering. Another possible research direction would be to consider finite approximations of the sinusoidal functions using classic CORDIC algorithms which build the approximation within a user-defined number of iterations [116]. Exploring implementations of these approximations in a hardware setting is crucial and we believe that an efficient on-chip sketch computation could lead to an overall reduction in logic resources compared to existing TCSPC histogram methods.

Another important consideration of implementing the sketched lidar approach on-chip is defining the level of precision used for computing the sketch. In Section 5.4.4 we constructed sketches of different size and various wordlengths and then analysed the quality of reconstruction. Interestingly, one could achieve good reconstructions by computing either a small sketch with high precision or a larger sketch with small precision. However, in this specific experiment, the dataset had a slightly large SBR of 6.82 and a moderate sized photon count. A future research direction would be to explore theoretically how different sketch sizes of various

wordlengths are effected by challenging conditions of low SBR and limited photon counts.

7.2.5 Complex Lidar Scenarios

In Part II, we assume that the background noise is constant over all time-intervals (i.e. $\mathcal{P}_b(x) = \frac{1}{T}$). However, in some practical settings the background may not be constant due to adverse weather conditions for example fog or snow. In Section 5.2.2, we developed sampling schemes that were essentially *blind* to uniform background noise. If we were to modify the lidar observation model in Eqn 5.1 to account for a non-constant background, then future work should explore if a different sampling scheme is needed to ensure that we sample in regions where the non-constant background is minimal.

A changing pulse width is another open-problem in the lidar community. The temporal response of a signal may not always be consistent due to the material or position of the target and the imaging conditions. Surfaces or objects are usually assumed to be opaque and approximately normal to the laser of the lidar device so that the reflected temporal response does not change across the captured scene. However, due to oblique angled surfaces, especially at long distances, the pulse of the signal may broaden and the shape of the pulse profile will vary across the measured pixels. Range-walk is a similar phenomena that occurs when capturing surfaces that are very reflective (e.g. retro-reflective materials). In this instance, the pulse width of the signal becomes extremely narrow (converges towards a Dirac delta function). In both scenarios, the effects of a changing pulse width can result in biased depth estimates of the scene. Although these problems are orthogonal to sketched lidar as a whole and affect existing methods to similar extents, a future line of research would tackle these problems and demonstrate that the sketched lidar framework is flexible to build these complexities into the true observation model. An immediate example of this is to compensate for broadening pulse width by replacing the constant IRF function h in Eqn 5.1 by a non-constant IRF h_b defined as

$$h_b(t) \propto \sum_{j=1}^{T} h(j) \exp\left(-\frac{(t-j)^2}{2(b-1)^2}\right) \tag{7.1}$$

where $b \in \mathbb{R}^+$ is an extra parameter that determines the broadening of pulse width. Notice that the non-constant IRF function is simply the existing constant IRF function convoluted with a Gaussian centred at 0 with standard deviation $b-1$. Due to the well-known convolution

theorem the resulting characteristic function of the signal model is simply

$$\Psi_{\mathcal{P}_s}(\omega) \;=\; \hat{h}(\omega)e^{-\frac{1}{2}(b-1)^2\omega^2}e^{\mathrm{i}\omega t}. \tag{7.2}$$

One can simply modify the sketched algorithms in Chapter 5 and 6 to incorporate the additional parameter b that is to be estimated.

7.2.6 Extensions to Other Photon Counting Imaging Modalities

Single photon counting techniques are not confined to single wavelength lidar and play an important role in other imaging domains. The closest relation is multi-spectral lidar imaging that gathers measurements of a scene at several spectral bands, making it possible to distinguish distinct materials in the field of view. In many existing multi-spectral approaches, a TCSPC histogram is obtained for each wavelength used. However, for many wavelengths, the memory and data transfer requirements quickly become infeasible using classic techniques. Recent approaches [144] attempt to integrate multiple wavelengths into a single histogram to reduce the overall complexities and size of the dataset acquired. A future line of research would be to develop a sketched multi-spectral method. Initially, one could construct a sketch for each wavelength and estimate the parameters of each spectral band in parallel. Another approach would be be to form a single sketch for all wavelengths, however one would need to explore if the decoder would be well-posed.

Aside from lidar, fluorescence-lifetime imaging microscopy (FLIM) is a well-established imaging technique in microscopy that attempts to distinguish the unique molecular environment of fluorophores to provide high-resolution images of living samples. A fluorophore that has been excited by a photon will drop to a ground state via a certain decay path quantified by a decay parameter τ. By recording the detected photons after excitement of a sample, we can estimate the specific decay rates of the fluorophore. Let x be the delay time of a photon, then a FLIM signal model [145] can be described by an exponential mixture model:

$$\mathcal{P}_s(x) \;=\; \sum_{k=1}^{K} I_0 e^{-\frac{x}{\tau_k}} \tag{7.3}$$

where I_0 is the intensity at time $t = 0$ and τ_k are the lifetimes of the K fluorophores. Similar to single-photon counting lidar, FLIM techniques also suffer from a data-transfer bottleneck. A

future direction of research would be to modify the framework set out in Chapter 5 to develop a compressive FLIM model. As the exponential distribution has a closed-form and well-behaved characteristic function, we could simply leverage the ECF sketches as used in Chapter 5. However, as the tails of the distributions provide important information on the decay rate of the fluorophores, it would be key to explore the role of random ECF sampling and the size of the sketches required to accurately estimate the decay rates.

References

[1] R. Gribonval, G. Blanchard, N. Keriven, and Y. Traonmilin, "Statistical learning guarantees for compressive clustering and compressive mixture modeling," *Mathematical Statistics and Learning*, vol. 3, no. 2, pp. 165–257, 2021.

[2] Y. Altmann, X. Ren, A. McCarthy, G. S. Buller, and S. McLaughlin, "Robust Bayesian target detection algorithm for depth imaging from sparse single-photon data," *IEEE Transactions on Computational Imaging*, vol. 2, no. 4, pp. 456–467, 2016.

[3] J. Tachella, Y. Altmann, N. Mellado, A. McCarthy, R. Tobin, G. S. Buller, J. Tourneret, and S. McLaughlin, "Real-time 3D reconstruction from single-photon lidar data using plug-and-play point cloud denoisers," *Nature communications*, vol. 10, no. 1, pp. 1–6, 2019.

[4] "Machine Learning CO2 Impact." `https://mlco2.github.io/impact/#co2eq`. Accessed: 01-12-2021.

[5] "OpenAI's massive GPT-3 model is impressive, but size isn't everything." `https://venturebeat.com/2020/06/01/ai-machine-learning-openai-gpt-3-size-isnt-everything/`. Accessed: 01-12-2021.

[6] R. Gribonval, G. Blanchard, N. Keriven, and Y. Traonmilin, "Compressive statistical learning with random feature moments," *Mathematical Statistics and Learning*, vol. 3, no. 2, pp. 113–164, 2021.

[7] V. Vapnik, "An overview of statistical learning theory," *IEEE Transactions on Neural Networks*, vol. 10, no. 5, pp. 988–999, 1999.

[8] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification, 2nd Edition*. Wiley, 2001.

[9] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer Series in Statistics, New York, NY, USA: Springer New York Inc., 2001.

[10] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[11] P. J. Bickel and C. A. Klaassen, *Efficient and adaptive estimation for semiparametric models*, vol. 4. Springer, 1998.

[12] M. R. Kosorok, *Introduction to empirical processes and semiparametric inference.* Springer, 2008.

[13] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4-5, pp. 411–430, 2000.

[14] A. J. Bell and T. J. Sejnowski, "An information-maximization approach to blind separation and blind deconvolution," *Neural Comput.*, vol. 7, p. 1129–1159, Nov. 1995.

[15] L. De Lathauwer, *Signal processing based on multilinear algebra*. Katholieke Universiteit Leuven Leuven, 1997.

[16] P. Comon, "Tensor diagonalization, a useful tool in signal processing," *IFAC Proceedings Volumes*, vol. 27, no. 8, pp. 77 – 82, 1994. IFAC Symposium on System Identification (SYSID'94), Copenhagen, Denmark, 4-6 July.

[17] P. Comon, "Tensor decompositions, state of the art and applications," *arXiv preprint arXiv:0905.0454*, 2009.

[18] L. De Lathauwer, B. De Moor, and J. Vandewalle, "An introduction to independent component analysis," *Journal of Chemometrics: A Journal of the Chemometrics Society*, vol. 14, no. 3, pp. 123–149, 2000.

[19] P. Comon, "Independent component analysis, a new concept?," *Signal processing*, vol. 36, no. 3, pp. 287–314, 1994.

[20] R. A. Fisher, "On the mathematical foundations of theoretical statistics," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 222, pp. 309–368, 1922.

[21] S.-i. A and H. Nagaoka, *Methods of information geometry*, vol. 191. American Mathematical Soc., 2007.

[22] L. P. Hansen, "Large sample properties of generalized method of moments estimators," *Econometrica*, vol. 50, no. 4, pp. 1029–1054, 1982.

[23] A. Hall, *Generalized Method of Moments*, pp. 230 – 255. Oxford University Press, 11 2007.

[24] B. O. Koopman, "On distributions admitting a sufficient statistic," *Transactions of the American Mathematical Society*, vol. 39, no. 3, pp. 399–409, 1936.

[25] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[26] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on information theory*, vol. 52, no. 2, pp. 489–509, 2006.

[27] E. J. Candès and M. B. Wakin, "An introduction to compressive sampling," *IEEE signal processing magazine*, vol. 25, no. 2, pp. 21–30, 2008.

[28] E. Candès and T. Tao, "Decoding by linear programming," *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[29] M. A. Davenport, M. F. Duarte, Y. C. Eldar, and G. Kutyniok, "Introduction to compressed sensing.," 2012.

[30] S. Foucart and H. Rauhut, "An invitation to compressive sensing," in *A mathematical introduction to compressive sensing*, pp. 1–39, Springer, 2013.

[31] Y. C. Eldar and G. Kutyniok, *Compressed sensing: theory and applications*. Cambridge university press, 2012.

[32] T. Blumensath, "Sampling and reconstructing signals from a union of linear subspaces," *IEEE Trans. Information Theory*, vol. 57, no. 7, pp. 4660–4671, 2011.

[33] T. Blumensath and M. E. Davies, "Iterative hard thresholding for compressed sensing," *Applied and computational harmonic analysis*, vol. 27, no. 3, pp. 265–274, 2009.

[34] T. Blumensath, M. E. Davies, and G. Rilling, "Greedy algorithms for compressed sensing," in *Compressed Sensing: Theory and Applications*, pp. 348–393, Cambridge University Press, 2012.

[35] E. J. Candès and Y. Plan, "Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements," *IEEE Transactions on Information Theory*, vol. 57, no. 4, pp. 2342–2359, 2011.

[36] R. G. Baraniuk and M. B. Wakin, "Random projections of smooth manifolds," *Foundations of computational mathematics*, vol. 9, no. 1, pp. 51–77, 2009.

[37] G. Puy, M. E. Davies, and R. Gribonval, "Recipes for stable linear embeddings from Hilbert spaces to $\mathbb{R}^m$," *IEEE Transactions on Information Theory*, vol. 63, no. 4, pp. 2171–2187, 2017.

[38] A. Bourrier, M. E. Davies, T. Peleg, P. Pérez, and R. Gribonval, "Fundamental performance limits for ideal decoders in high-dimensional linear inverse problems," *IEEE Transactions on Information Theory*, vol. 60, no. 12, pp. 7928–7946, 2014.

[39] R. Gribonval, A. Chatalic, N. Keriven, V. Schellekens, L. Jacques, and P. Schniter, "Sketching data sets for large-scale learning: Keeping only what you need," *IEEE Signal Processing Magazine*, vol. 38, no. 5, pp. 12–36, 2021.

[40] G. Cormode, M. Garofalakis, P. J. Haas, and C. Jermaine, "Synopses for massive data: Samples, histograms, wavelets, sketches," *Foundations and Trends in Databases*, vol. 4, no. 1–3, pp. 1–294, 2012.

[41] N. Aronszajn, "Theory of reproducing kernels," *Transactions of the American Mathematical Society*, vol. 68, no. 3, p. 337–337, 1950.

[42] T. Hofmann, B. Schölkopf, and A. J. Smola, "Kernel methods in machine learning," *The annals of statistics*, vol. 36, no. 3, pp. 1171–1220, 2008.

[43] S. Bochner, "Monotone funktionen, stieltjessche integrale und harmonische analyse," *Mathematische Annalen*, vol. 108, no. 1, p. 378–410, 1933.

[44] K. Muandet, K. Fukumizu, B. Sriperumbudur, B. Schölkopf, *et al.*, "Kernel mean embedding of distributions: A review and beyond," *Foundations and Trends® in Machine Learning*, vol. 10, no. 1-2, pp. 1–141, 2017.

[45] N. Keriven, *Sketching for large-scale learning of mixture models*. Theses, Université Rennes 1, Oct. 2017.

[46] N. Keriven, A. Bourrier, R. Gribonval, and P. Pérez, "Sketching for large-scale learning of mixture models," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pp. 6190–6194, IEEE, 2016.

[47] A. Chatalic, V. Schellekens, F. Houssiau, Y.-A. de Montjoye, L. Jacques, and R. Gribonval, "Compressive learning with privacy guarantees," *Information and Inference*, 2021.

[48] D. Feldman, M. Schmidt, and C. Sohler, "Turning big data into tiny data: Constant-size coresets for k-means, pca, and projective clustering," *SIAM Journal on Computing*, vol. 49, no. 3, pp. 601–657, 2020.

[49] C. Williams and M. Seeger, "Using the Nyström method to speed up kernel machines," in *Proceedings of the 14th annual conference on neural information processing systems*, no. CONF, pp. 682–688, 2001.

[50] P. Drineas, M. W. Mahoney, and N. Cristianini, "On the Nyström method for approximating a gram matrix for improved kernel-based learning.," *journal of machine learning research*, vol. 6, no. 12, 2005.

[51] W. B. Johnson and J. Lindenstrauss, "Extensions of Lipschitz mappings into a Hilbert space 26," *Contemporary mathematics*, vol. 26, 1984.

[52] D. Achlioptas, "Database-friendly random projections," in *Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pp. 274–281, ACM, 2001.

[53] N. Ailon and B. Chazelle, "Approximate nearest neighbors and the fast Johnson-Lindenstrauss transform," in *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*, pp. 557–563, 2006.

[54] J. A. Lee and M. Verleysen, *Nonlinear dimensionality reduction*. Springer Science & Business Media, 2007.

[55] G. Cormode and S. Muthukrishnanb, "An improved data stream summary: the count-min sketch and its applications," *Journal of Algorithms*, vol. 55, pp. 58–75, 2005.

[56] J. A. Tropp, A. Yurtsever, M. Udell, and V. Cevher, "Practical sketching algorithms for low-rank matrix approximation," *SIAM Journal on Matrix Analysis and Applications*, vol. 38, no. 4, pp. 1454–1485, 2017.

[57] J. A. Tropp, A. Yurtsever, M. Udell, and V. Cevher, "Streaming low-rank matrix approximation with an application to scientific simulation," *SIAM Journal on Scientific Computing*, vol. 41, no. 4, pp. A2430–A2463, 2019.

[58] J. Gao, J. Sun, J. Wei, and Q. Wang, "Research of underwater target detection using a slit streak tube imaging lidar," in *2011 Academic International Symposium on Optoelectronics and Microelectronics Technology*, pp. 240–243, 2011.

[59] M. Pierzchała, P. Giguère, and R. Astrup, "Mapping forests using an unmanned ground vehicle with 3D lidar and graph-slam," *Computers and Electronics in Agriculture*, vol. 145, pp. 217 – 225, 2018.

[60] J. Hecht, "Lidar for self-driving cars," *Opt. Photon. News*, vol. 29, pp. 26–33, Jan 2018.

[61] J. Rapp, J. Tachella, Y. Altmann, S. McLaughlin, and V. K. Goyal, "Advances in single-photon lidar for autonomous vehicles: Working principles, challenges, and recent advances," *IEEE Signal Processing Magazine*, vol. 37, no. 4, pp. 62–71, 2020.

[62] S. Hernandez-Marin, A. M. Wallace, and G. J. Gibson, "Bayesian analysis of lidar signals with multiple returns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2170–2180, 2007.

[63] Y. Altmann, X. Ren, A. McCarthy, G. S. Buller, and S. McLaughlin, "Lidar waveform-based analysis of depth images constructed using sparse single-photon data," *IEEE Transactions on Image Processing*, vol. 25, no. 5, pp. 1935–1946, 2016.

[64] Y. Altmann and S. McLaughlin, "Range estimation from single-photon lidar data using a stochastic EM approach," in *2018 26th European Signal Processing Conference (EUSIPCO)*, pp. 1112–1116, 2018.

[65] R. K. Henderson, N. Johnston, H. Chen, D. D. Li, G. Hungerford, R. Hirsch, D. McLoskey, P. Yip, and D. J. S. Birch, "A 192×128 time correlated single photon counting imager in 40nm CMOS technology," in *ESSCIRC 2018 - IEEE 44th European Solid State Circuits Conference (ESSCIRC)*, pp. 54–57, 2018.

[66] X. Ren, P. W. R. Connolly, A. Halimi, Y. Altmann, S. McLaughlin, I. Gyongy, R. K. Henderson, and G. S. Buller, "High-resolution depth profiling using a range-gated CMOS SPAD quanta image sensor," *Opt. Express*, vol. 26, pp. 5541–5557, Mar 2018.

[67] R. J. Walker, J. A. Richardson, and R. K. Henderson, "A 128×96 pixel event-driven phase-domain $\Delta\Sigma$-based fully digital 3D camera in 0.13μm CMOS imaging technology," in *2011 IEEE International Solid-State Circuits Conference*, pp. 410–412, 2011.

[68] F. M. Della Rocca, H. Mai, S. W. Hutchings, T. Al Abbas, A. Tsiamis, P. Lomax, I. Gyongy, N. A. W. Dutton, and R. K. Henderson, "A 128 × 128 SPAD dynamic vision-triggered time of flight imager," in *ESSCIRC 2019 - IEEE 45th European Solid State Circuits Conference (ESSCIRC)*, pp. 93–96, 2019.

[69] F. Mattioli Della Rocca, H. Mai, S. W. Hutchings, T. A. Abbas, K. Buckbee, A. Tsiamis, P. Lomax, I. Gyongy, N. A. W. Dutton, and R. K. Henderson, "A 128 × 128 SPAD motion-triggered time-of-flight image sensor with in-pixel histogram and column-parallel vision processor," *IEEE Journal of Solid-State Circuits*, vol. 55, no. 7, pp. 1762–1775, 2020.

[70] S. W. Hutchings, N. Johnston, I. Gyongy, T. Al Abbas, N. A. W. Dutton, M. Tyler, S. Chan, J. Leach, and R. K. Henderson, "A reconfigurable 3-D-stacked SPAD imager with in-pixel histogramming for flash lidar or high-speed time-of-flight imaging," *IEEE Journal of Solid-State Circuits*, vol. 54, no. 11, pp. 2947–2956, 2019.

[71] C. Zhang, S. Lindner, I. M. Antolović, J. Mata Pavia, M. Wolf, and E. Charbon, "A 30-frames/s, 252×144 SPAD flash lidar with 1728 dual-clock 48.8-ps tdcs, and pixel-wise integrated histogramming," *IEEE Journal of Solid-State Circuits*, vol. 54, no. 4, pp. 1137–1151, 2019.

[72] J. Rapp, M. A. Dawson, R., and V. K. Goyal, "Dithered depth imaging," *Optics Express*, vol. 28, no. 23, pp. 35143–35157, 2020.

[73] A. Kadambi and P. T. Boufounos, "Coded aperture compressive 3-D lidar," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1166–1170, 2015.

[74] A. Halimi, P. Ciuciu, A. Mccarthy, S. Mclaughlin, and G. S. Buller, "Fast adaptive scene sampling for single-photon 3D lidar images," in *IEEE CAMSAP 2019 - International Workshop on Computational Advances in Multi-Sensor Adaptive Processing*, (Le Gosier (Guadeloupe), France), Dec. 2019.

[75] I. Maksymova, C. Steger, and N. Druml, "Review of lidar sensor data acquisition and compression for automotive applications," in *Multidisciplinary Digital Publishing Institute Proceedings*, vol. 2, p. 852, 2018.

[76] M. P. Sheehan, M. S. Kotzagiannidis, and M. E. Davies, "Compressive independent component analysis," in *2019 27th European Signal Processing Conference (EUSIPCO)*, pp. 1–5, 2019.

[77] M. P. Sheehan and M. E. Davies, "Compressive independent component analysis: Theory and algorithms," *To Appear in Information and Inference: A Journal of the IMA*, 2022.

[78] S. Szarek, "Nets of Grassmann manifold and orthogonal group," in *Proceedings of Research Workshop on Banach Space Theory*, pp. 169–186, 06 1981.

[79] F. Krahmer and R. Ward, "New and improved Johnson-Lindenstrauss embeddings via the restricted isometry property," *SIAM Journal on Mathematical Analysis*, vol. 43, no. 3, pp. 1269–1281, 2011.

[80] J. A. Tropp, "Improved analysis of the subsampled randomized Hadamard transform," *Advances in Adaptive Data Analysis*, vol. 3, no. 1, pp. 115–126, 2011.

[81] H. Rauhut, R. Schneider, and Ž. Stojanac, "Low rank tensor recovery via iterative hard thresholding," *Linear Algebra and its Applications*, vol. 523, pp. 220–262, 2017.

[82] R. Vershynin, "Introduction to the non-asymptotic analysis of random matrices," in *Compressed Sensing: Theory and Applications*, pp. 210–268, Cambridge University Press, 2012.

[83] J. F. Cardoso, B. L. R. De Moor, and M. S. Moonen, "A tetradic decomposition of 4th-order tensors : Application to the source separation problem," (Amsterdam), pp. 375–382, Elsevier;, 1995.

[84] J. F. Cardoso, "Fourth-order cumulant structure forcing: application to blind array processing," in *[1992] IEEE Sixth SP Workshop on Statistical Signal and Array Processing*, pp. 136–139, 1992.

[85] J. Cadzow, "Signal enhancement-a composite property mapping algorithm," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 36, pp. 49–62, 1988.

[86] Z. Wen and W. Yin, "A feasible method for optimization with orthogonality constraints," *Math. Program.*, vol. 142, p. 397–434, Dec. 2013.

[87] J. Crank and P. Nicolson, "A practical method for numerical evaluation of solutions of partial differential equations of the heat-conduction type," in *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 43, pp. 50–67, Cambridge University Press, 1947.

[88] F. Woolfe, E. Liberty, V. Rokhlin, and M. Tygert, "A fast randomized algorithm for the approximation of matrices," *Applied and Computational Harmonic Analysis*, vol. 25, no. 3, pp. 335 – 366, 2008.

[89] J. F. Cardoso and A. Souloumiac, "Blind beamforming for non-gaussian signals," in *IEE proceedings F (radar and signal processing)*, vol. 140, pp. 362–370, IET, 1993.

[90] E. Massart and V. Abrol, "Coordinate descent on the orthogonal group for recurrent neural network training," *arXiv preprint arXiv:2108.00051*, 2021.

[91] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp, "Living on the edge: phase transitions in convex programs with random data," *Information and Inference: A Journal of the IMA*, vol. 3, no. 3, pp. 224–294, 2014.

[92] S. Amari, A. Cichocki, and H. Yang, "A new learning algorithm for blind signal separation," in *Advances in neural information processing systems*, pp. 757–763, 1996.

[93] F. Bach and M. Jordan, "Kernel independent component analysis," *Journal of machine learning research*, vol. 3, no. Jul, pp. 1–48, 2002.

[94] J. Higham, W. Brevis, and C. Keylock, "Implications of the selection of a particular modal decomposition technique for the analysis of shallow flows," *Journal of Hydraulic Research*, vol. 56, no. 6, pp. 796–805, 2018.

[95] W. Brevis and M. García-Villalba, "Shallow-flow visualization analysis by proper orthogonal decomposition," *Journal of Hydraulic Research*, vol. 49, no. 5, pp. 586–594, 2011.

[96] A. Hyvarinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Transactions on Neural Networks*, vol. 10, no. 3, pp. 626–634, 1999.

[97] K. L. Clarkson, "Tighter bounds for random projections of manifolds," in *Proceedings of the twenty-fourth annual symposium on Computational geometry*, pp. 39–48, 2008.

[98] T. G. Kolda and B. W. Bader, "Tensor decompositions and applications," *SIAM review*, vol. 51, no. 3, pp. 455–500, 2009.

[99] R. Coleman, *Calculus on Normed Vector Spaces*. Springer New York, 2012.

[100] A. Rahimi and B. Recht, "Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning," in *Advances in neural information processing systems*, pp. 1313–1320, 2009.

[101] A. Cohen, W. Dahmen, and R. DeVore, "Compressed sensing and best k-term approximation," *Journal of the American mathematical society*, vol. 22, no. 1, pp. 211–231, 2009.

[102] M. P. Sheehan, A. Gonon, and M. E. Davies, "Compressive learning for semi-parametric models," *arXiv preprint arXiv:1910.10024*, 2019.

[103] R. Vidal, Y. Ma, and S. Sastry, "Generalized principal component analysis (gpca)," *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 12, pp. 1945–1959, 2005.

[104] R. Vidal, "Subspace clustering," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 52–68, 2011.

[105] H. Derksen, "Hilbert series of subspace arrangements," *Journal of pure and applied algebra*, vol. 209, no. 1, pp. 91–98, 2007.

[106] Y. Ma, A. Y. Yang, H. Derksen, and R. Fossum, "Estimation of subspace arrangements with applications in modeling and segmenting mixed data," *SIAM review*, vol. 50, no. 3, pp. 413–458, 2008.

[107] R. Vidal, "A tutorial on subspace clustering," 2010.

[108] R. Kueng, H. Rauhut, and U. Terstiege, "Low rank matrix recovery from rank one measurements," *Applied and Computational Harmonic Analysis*, vol. 42, no. 1, pp. 88–116, 2017.

[109] M. A. Davenport and J. Romberg, "An overview of low-rank matrix recovery from incomplete observations," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 4, pp. 608–622, 2016.

[110] M. P. Sheehan, J. Tachella, and M. E. Davies, "A sketching framework for reduced data transfer in photon counting lidar," *IEEE Transactions on Computational Imaging*, vol. 7, pp. 989–1004, 2021.

[111] S. R. Jammalamadaka and A. Sengupta, *Topics in circular statistics*, vol. 5. world scientific, 2001.

[112] P. Padmanabhan, C. Zhang, and E. Charbon, "Modeling and analysis of a direct time-of-flight sensor architecture for LiDAR applications," *Sensors*, vol. 19, no. 24, 2019.

[113] A. Bashirov, "Chapter 12 - Fourier series and integrals," in *Mathematical Analysis Fundamentals*, pp. 307–345, Boston: Elsevier, 2014.

[114] N. Keriven, A. Bourrier, R. Gribonval, and P. Pérez, "Sketching for large-scale learning of mixture models," *Information and Inference: A Journal of the IMA*, vol. 7, no. 3, pp. 447–508, 2018.

[115] N. Keriven, D. Garreau, and I. Poli, "Newma: a new method for scalable model-free online change-point detection," *IEEE Transactions on Signal Processing*, vol. 68, pp. 3515–3528, 2020.

[116] F. de Dinechin, M. Istoan, and G. Sergent, "Fixed-point trigonometric functions on FPGAs," *SIGARCH Comput. Archit. News*, vol. 41, p. 83–88, June 2014.

[117] V. Schellekens and L. Jacques, "Asymmetric compressive learning guarantees with applications to quantized sketches," *IEEE Transactions on Signal Processing*, vol. 70, pp. 1348–1360, 2022.

[118] A. Feuerverger and P. McDunnough, "On the efficiency of empirical characteristic function procedures," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 43, no. 1, pp. 20–27, 1981.

[119] A. Feuerverger and P. McDunnough, "On some Fourier methods for inference," *Journal of the American Statistical Association*, vol. 76, no. 374, pp. 379–387, 1981.

[120] L. P. Hansen, J. Heaton, and A. Yaron, "Finite-sample properties of some alternative GMM estimators," *Journal of Business & Economic Statistics*, vol. 14, no. 3, pp. 262–280, 1996.

[121] J. Hausman, R. Lewis, K. Menzel, and W. Newey, "Properties of the CUE estimator and a modification with moments," *Journal of Econometrics*, vol. 165, no. 1, pp. 45 – 57, 2011. Moment Restriction-Based Econometric Methods.

[122] D. Shin, A. Kirmani, V. K. Goyal, and J. H. Shapiro, "Photon-efficient computational 3-D and reflectivity imaging with single-photon detectors," *IEEE Transactions on Computational Imaging*, vol. 1, no. 2, pp. 112–125, 2015.

[123] J. Rapp and V. Goyal, "A few photons among many: Unmixing signal and noise for photon-efficient active imaging," *IEEE Transactions on Computational Imaging*, vol. PP, 09 2016.

[124] G. Turin, "An introduction to matched filters," *IRE Transactions on Information Theory*, vol. 6, no. 3, pp. 311–329, 1960.

[125] R. M. Marino and W. R. Davis, "Jigsaw: a foliage-penetrating 3D imaging laser radar system," *Lincoln Lab J.*, vol. 15, no. 1, pp. 23–36, 2005.

[126] I. Gyongy, S. W. Hutchings, A. Halimi, M. Tyler, S. Chan, F. Zhu, S. McLaughlin, R. K. Henderson, and J. Leach, "High-speed 3D sensing via hybrid-mode imaging and guided upsampling," *Optica*, vol. 7, pp. 1253–1260, Oct 2020.

[127] J. Tachella, Y. Altmann, X. Ren, A. McCarthy, G. S. Buller, S. McLaughlin, and J. Tourneret, "Bayesian 3D reconstruction of complex scenes from single-photon lidar data," *SIAM Journal on Imaging Sciences*, vol. 12, pp. 521–550, 03 2019.

[128] A. Halimi, R. Tobin, A. McCarthy, S. McLaughlin, and G. S. Buller, "Restoration of multilayered single-photon 3D lidar images," in *2017 25th European Signal Processing Conference (EUSIPCO)*, pp. 708–712, 2017.

[129] R. Tobin, A. Halimi, A. McCarthy, X. Ren, K. J. McEwan, S. McLaughlin, and G. S. Buller, "Long-range depth profiling of camouflaged targets using single-photon detection," *Optical Engineering*, vol. 57, no. 3, p. 031303, 2017.

[130] M. P. Sheehan, J. Tachella, and M. E. Davies, "Surface detection for sketched single photon lidar," in *2021 29th European Signal Processing Conference (EUSIPCO)*, pp. 621–625, 2021.

[131] J. Tachella, M. P. Sheehan, and M. E. Davies, "Sketched RT3D: How to reconstruct billions of photons per second," in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022.

[132] J. Tachella, Y. Altmann, S. McLaughlin, and J. . Y. Tourneret, "Fast surface detection in single-photon lidar waveforms," in *2019 27th European Signal Processing Conference (EUSIPCO)*, pp. 1–5, 2019.

[133] D. B. Campbell and C. A. Oprian, "On the kolmogorov-smirnov test for the poisson distribution with unknown mean," *Biometrical Journal*, vol. 21, no. 1, pp. 17–24, 1979.

[134] Y. Fan, "Goodness-of-Fit Tests for a Multivariate Distribution by the Empirical Characteristic Function," *Journal of Multivariate Analysis*, vol. 62, pp. 36–63, July 1997.

[135] I. A. Koutrouvelis and J. Kellermeier, "A goodness-of-fit test based on the empirical characteristic function when parameters must be estimated," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 43, no. 2, pp. 173–176, 1981.

[136] J. Rapp, Y. Ma, R. M. Dawson, and V. K. Goyal, "High-flux single-photon lidar," *Optica*, vol. 8, no. 1, pp. 30–39, 2021.

[137] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 723–773, 2012.

[138] D. Shin, A. Kirmani, V. K. Goyal, and J. H. Shapiro, "Photon-efficient computational 3-D and reflectivity imaging with single-photon detectors," *IEEE Trans. Comput. Imaging*, vol. 1, no. 2, pp. 112–125, 2015.

[139] Y. Altmann, X. Ren, A. McCarthy, G. S. Buller, and S. McLaughlin, "Target detection for depth imaging using sparse single-photon data," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3256–3260, March 2016.

[140] D. Shin, F. Xu, F. N. Wong, J. H. Shapiro, and V. K. Goyal, "Computational multi-depth single-photon imaging," *Optics express*, vol. 24, no. 3, pp. 1873–1888, 2016.

[141] D. B. Lindell, M. O'Toole, and G. Wetzstein, "Single-Photon 3D Imaging with Deep Sensor Fusion," *ACM Trans. Graph. (SIGGRAPH)*, no. 4, 2018.

[142] S. V. Venkatakrishnan, C. A. Bouman, and B. Wohlberg, "Plug-and-play priors for model based reconstruction," in *2013 IEEE Global Conference on Signal and Information Processing*, pp. 945–948, 2013.

[143] A. McCarthy, R. J. Collins, N. J. Krichel, V. Fernández, A. M. Wallace, and G. S. Buller, "Long-range time-of-flight scanning sensor based on high-speed time-correlated single-photon counting," *Appl. Opt.*, vol. 48, pp. 6241–6251, Nov 2009.

[144] X. Ren, Y. Altmann, R. Tobin, A. Mccarthy, S. Mclaughlin, and G. S. Buller, "Wavelength-time coding for multispectral 3d imaging using single-photon lidar," *Opt. Express*, vol. 26, pp. 30146–30161, Nov 2018.

[145] R. Datta, T. M. Heaster, J. T. Sharick, A. A. Gillette, and M. C. Skala, "Fluorescence lifetime imaging microscopy: fundamentals and advances in instrumentation, analysis, and applications," *Journal of Biomedical Optics*, vol. 25, no. 7, pp. 1 – 43, 2020.