From Bits to Information:
Learning meets Compressive Sensing

by

Youssef Mroueh

École Polytechnique (2010)
École des Mines de Paris (2011)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

February 2015

© Massachusetts Institute of Technology 2015. All rights reserved.
From Bits to Information: 
Learning meets Compressive Sensing 
by 
Youssef Mroueh 

Submitted to the Department of Electrical Engineering and Computer Science 
on January 2015, in partial fulfillment of the 
requirements for the degree of 
Doctor of Philosophy 

Abstract 
A quantization approach to supervised learning, compressive sensing, and phase retrieval is presented in this thesis. We introduce a set of common techniques that allow us, in those three settings, to represent high dimensional data using the order statistics of linear and non linear measurements. We introduce new algorithms for signals classification in the multiclass and the multimodal settings, as well as algorithms for signals representation and recovery from quantized linear and quadratic measurements. We analyze the statistical consistency of our algorithms and prove their robustness to different sources of perturbation, as well as their computational efficiency. We present and analyze applications of our theoretical results in realistic setups, such as computer vision classification tasks, Audio-Visual Automatic Speech Recognition, lossy image compression and retrieval via locality sensitive hashing, locally linear estimation in large scale learning and Fourier sampling for phase retrieval - of particular interest in X-ray crystallography and super-resolution diffraction imaging applications. Our analysis of quantization based algorithms highlights interesting tradeoffs between memory complexity, sample complexity, and time complexity in algorithms design. 

Thesis Supervisor: Professor Tomaso Poggio 
Title: Eugene McDermott Professor of Brain and Cognitive Sciences 

Thesis Supervisor: Dr Lorenzo Rosasco 
Title: Visiting Faculty
Acknowledgments

This thesis is dedicated to the memory of my father, and to my mother, that I start by acknowledging for a lifetime support and unconditional sacrifices. To my sisters Lina, and Malak, and to my brother Kassem, whose help, support, love and care, accompanied and cherished me, all along my studies.

Foremost, I would like to thank my advisor Tommy Poggio for offering me the unique opportunity of joining his group, first as visiting student and then as a PhD student. Tommy maintains an exciting and an eclectic research environment, his ideas, his energy, his determination and his passion for synergies between diverse research fields, inspired me throughout my thesis. My gratitude goes as well to my co-advisor Lorenzo Rosasco, for believing in me, for his continuous support, and for giving me the freedom in exploring my ideas and in finding my way in research. I wish also to thank Lorenzo for our collaboration and for his warm and generous hospitality during my visits to the university of Genoa, and the Italian Institute of Technology.

I am also thankful to the members of my thesis committee: Piotr Indyk who initiated me to the beauty of randomization in algorithms and in modern signal processing, his class on Geometric Computing inspired the second part of this thesis, on compressive Sensing. Bill Freeman, whose work in computer vision and machine learning has been always a source of inspiration for me. Laurent Demanet, who was always available for discussing my work, his feedback has been valuable in various stages of my PhD.

I was also fortunate to work and interact with many brilliant people who shaped my mind and thinking. I would particularly like to thank Stéphane Mallat, Stéphane Gaubert and Joan Bruna, for introducing me to the world of signal processing, optimization and machine learning in École polytechnique; Jean-Jacques Slotine, for our collaboration and for hosting me as a visiting student in his group at MIT; Brigitte D’Andrea-Novel for her support in Ecole des Mines de Paris and for encouraging me to take the PhD path; Ramesh Raskar and Gordon Wetzstein, from the Camera Culture Group at MIT, for their class on computational photography, and the discussions we had, which inspired the third part of this thesis on Phase Retrieval.

My deep gratitude goes also to my great mentors, and collaborators over my summer internship at IBM Watson, Vaibahava Goel, Etienne Marcheret and Steven Rennie. I learned a lot from this experience on Audio-Visual Automatic Speech Recognition and was fortunate to be able to push over that internship other aspects of my thesis, that culminates in two chapters of this thesis (Chapters 6 and 8). I wish also to thank Guillaume Lecué who kindly suggested pointers on empirical processes, which were useful in the course of this thesis.
This thesis would not have been possible without the support of Gadi Geiger. At various difficult moments, Gadi’s advices, coffee, pool games and apple tarts were always a breath of fresh air. I wish also to thank Kathleen Sullivan, and Felecia Bishop who were the key persons in overcoming all the administrative complications. In particular Kathleen was always supportive and helpful, also thanks Kathleen for introducing me to the delicious Bibimbap.

I am also fortunate to have amazing and supportive friends. In particular Edouard Verdier, Steve Voinea, and George Zahariade. Edouard, having you as a flatmate at MIT was a lucky hazard that added a brother to my family. Steve you are a dear and wonderful friend for a lifetime, my bostonian journey would not have been the same without your unconditional support, your positive energy and your refined food taste. George, you have been always a lifelong friend that I can count on, and that I can bother with my research and with any problem. Your input has been valuable, as you discussed and challenged me, on most of the contents of this thesis.

I wish also to thank my friends in Poggio Lab, past and present, Leyla Isik, Andrea Tacchetti, Charlie Frogner, Chiyuan Zhang, Owen Lewis, Jim Much, Ethan Meyers, Fabio Anselmi, Georgios Evangelopoulos, Guille De Canas, Yuzhao (Allen) Ni, Chun-Kai Wang, Stav Braun, Nicholas Edelman, Pavan Mallapragada, Carlo Ciliberto, Max Nickel, Alessandro Rudi, Silvia Villa, Matteo Santoro, Jake Bouvrie, Joel Leibo and Qianli liao, for maintaining a fun and stimulating work environment. Last but not least special thanks to Ludwig Schmidt for stimulating discussions on various topics in theoretical computer sciences.
## Contents

1 From Bits to Information .................................................. 13
   1.1 Problem Area and State of the Field .............................. 14
   1.2 Thesis Contributions ................................................ 17

I Multiclass & Multimodal Learning .................................. 19

2 Multiclass Learning via Simplex Coding .......................... 21
   2.1 Introduction .......................................................... 21
   2.2 Background ........................................................... 22
      2.2.1 Relaxation Approach to Binary Classification ............ 23
      2.2.2 Relaxation Error Analysis for Binary Classification ...... 24
      2.2.3 Previous Work on Multiclass Classification ............... 25
   2.3 A Relaxation Approach to Multi-Class Classification ......... 26
      2.3.1 Simplex Coding .................................................. 26
      2.3.2 A Geometric Relaxation approach ............................. 27
      2.3.3 Multiclass Margin ................................................ 27
      2.3.4 Multiclass Simplex Loss Functions ........................... 27
      2.3.5 Expected V-Risk Minimization ................................. 30
   2.4 Conclusion ............................................................ 30

3 Consistency Analysis of Simplex Multiclass Learning .......... 31
   3.1 Introduction .......................................................... 31
   3.2 Classification Calibration .......................................... 33
      3.2.1 Examples ......................................................... 36
   3.3 Conclusion ............................................................ 36
   3.4 Proofs of Lemmas 1, and 2 ......................................... 36

4 Multiclass Classification with Simplex Support Vector Machines 40
   4.1 Introduction .......................................................... 40
   4.2 Multiclass SVM: Previous Work .................................... 42
   4.3 DC-SVM Calibration and Comparison Inequality ................. 42
4.4 Regularization in Vector Valued RKHS

<table>
<thead>
<tr>
<th>4.4.1 Vector Valued RKHS (vv-RKHS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.4.2 Generalization Bounds</td>
</tr>
</tbody>
</table>

4.5 Algorithms

| 4.5.1 Dual Formulation              |
| 4.5.2 Primal Solution with Stochastic Gradient |
| 4.5.3 Computing the Simplex Coding   |

4.6 Conclusion

4.7 Proofs of Theorems 2 and 3 and Proposition 1

5 Multiclass Classification with Simplex Least Squares

| 5.1 Introduction                     |
| 5.2 SLS Calibration and Comparison Inequality |
| 5.2.1 Improved Exponents under Noise Condition |
| 5.3 Regularized SLS in vv-RKHS        |
| 5.3.1 Regularized SLS                 |
| 5.3.2 Generalization Bounds           |
| 5.4 Algorithms                        |
| 5.4.1 Regularization Path and Model Selection |
| 5.4.2 Online/Incremental Optimization |
| 5.5 Numerical Validation for Classification via Simplex Coding |
| 5.5.1 Comparison of Computational Complexity |
| 5.5.2 Numerical Results               |
| 5.6 Proofs of Theorem 4 and Theorem 5 |

6 Deep Multimodal Learning

| 6.1 Introduction                      |
| 6.2 Audio-Visual Data Set & Feature Extraction |
| 6.2.1 IBM AV-ASR Large Vocabulary Studio Dataset |
| 6.2.2 Feature Extraction              |
| 6.2.3 Context-dependent Phoneme Targets |
| 6.3 Uni-modal DNNs & Feature Fusion   |
| 6.4 Bilinear Deep Neural Network      |
| 6.4.1 Factored Bilinear Softmax       |
| 6.4.2 Factored Bilinear Softmax With Sharing |
| 6.5 Back-propagation with the Factored Bilinear DNN with Sharing |
| 6.6 Combining Posteriors from Bimodal and Bilinear Bimodal Networks |
| 6.7 Conclusion                       |
# II Q-ary Compressive Sensing and Embedding

## 7 Q-ary Compressive Sensing

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.1</td>
<td>Introduction</td>
<td>80</td>
</tr>
<tr>
<td>7.2</td>
<td>Q-ary Compressive Sensing</td>
<td>83</td>
</tr>
<tr>
<td>7.2.1</td>
<td>Sensing and Recovery</td>
<td>83</td>
</tr>
<tr>
<td>7.2.2</td>
<td>The signal Set &amp; the Gaussian Mean Width</td>
<td>85</td>
</tr>
<tr>
<td>7.3</td>
<td>Recovery guarantees</td>
<td>86</td>
</tr>
<tr>
<td>7.4</td>
<td>Experimental Validation</td>
<td>87</td>
</tr>
<tr>
<td>7.4.1</td>
<td>An Algorithm for Sparse recovery</td>
<td>87</td>
</tr>
<tr>
<td>7.4.2</td>
<td>Sparse Recovery</td>
<td>87</td>
</tr>
<tr>
<td>7.4.3</td>
<td>Image Reconstruction</td>
<td>88</td>
</tr>
<tr>
<td>7.5</td>
<td>Conclusion</td>
<td>89</td>
</tr>
<tr>
<td>7.6</td>
<td>Theoretical Analysis</td>
<td>90</td>
</tr>
</tbody>
</table>

## 8 Random Max-Out Embedding

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.1</td>
<td>Introduction</td>
<td>95</td>
</tr>
<tr>
<td>8.2</td>
<td>Deep Max-Out Networks</td>
<td>97</td>
</tr>
<tr>
<td>8.3</td>
<td>Random Max-Out Features</td>
<td>97</td>
</tr>
<tr>
<td>8.3.1</td>
<td>Discussion of the Derived Max-Out Kernel</td>
<td>98</td>
</tr>
<tr>
<td>8.3.2</td>
<td>Locality Sensitive Hashing</td>
<td>100</td>
</tr>
<tr>
<td>8.3.3</td>
<td>Sample Complexity of the Random Max-Out Feature Map</td>
<td>101</td>
</tr>
<tr>
<td>8.4</td>
<td>Numerical Experiments</td>
<td>101</td>
</tr>
<tr>
<td>8.5</td>
<td>Proofs of Propositions 2 and 3</td>
<td>102</td>
</tr>
<tr>
<td>8.5.1</td>
<td>Technical tools: Covering and Bracketing Numbers</td>
<td>104</td>
</tr>
</tbody>
</table>

# III Phase Retrieval

## 9 One Bit Phase Retrieval

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.1</td>
<td>Introduction</td>
<td>110</td>
</tr>
<tr>
<td>9.2</td>
<td>Background and Previous Work</td>
<td>112</td>
</tr>
<tr>
<td>9.3</td>
<td>Summary of Our Main Results and Techniques</td>
<td>116</td>
</tr>
<tr>
<td>9.3.1</td>
<td>Phase Recovery from Severely Perturbed Measurements</td>
<td>116</td>
</tr>
<tr>
<td>9.3.2</td>
<td>One-Bit Phase Retrieval and Alternating Minimization</td>
<td>117</td>
</tr>
<tr>
<td>9.4</td>
<td>One-Bit Phase Retrieval</td>
<td>118</td>
</tr>
<tr>
<td>9.4.1</td>
<td>Quantization and Recovery</td>
<td>118</td>
</tr>
<tr>
<td>9.4.2</td>
<td>Geometric Intuition</td>
<td>119</td>
</tr>
<tr>
<td>9.4.3</td>
<td>One bit Phase Retrieval from Distorted and Noisy measurements</td>
<td>120</td>
</tr>
<tr>
<td>9.4.4</td>
<td>Main Results for One Bit Phase Retrieval</td>
<td>122</td>
</tr>
<tr>
<td>9.5</td>
<td>Theoretical Analysis</td>
<td>123</td>
</tr>
<tr>
<td>9.5.1</td>
<td>One Bit Phase Retrieval: Correctness and Concentration</td>
<td>124</td>
</tr>
</tbody>
</table>
List of Figures

1-1 A unimodal multi-class problem consists in learning a model that predicts the semantic class label of a unimodal signal. .................................................. 14
1-2 An example of multimodal multi-class problems is the Audio-Visual Automatic Speech Recognition problem (AV-ASR). .................................................. 15
1-3 Compressive acquisition of a scene on the alphabet \{0, 1, 2, 3\} .......... 15
1-4 A redundant representation defined on a finite alphabet in a x-ray diffraction imaging. ............................................................................................................. 16

2-1 Surrogate losses in binary classification. .................................................. 23
2-2 Decoding with simplex coding \(q = 3\). .................................................. 26
2-3 Different geometric margin definitions. While in HS points in the same class are mapped to a half space, in LS they are mapped to a sphere. In DC they are mapped to a cone, and in AC to a voronoi cell of the corresponding vertex. 28
2-4 Level sets of different simplex losses: Discriminative Conic Margin. .... 29
2-5 Level sets of different losses: All pairs Conic Margin. ......................... 29

6-1 Low Level Features Extraction for the Audio and the Visual Channels. . 70
6-2 Bimodal DNN. .......................................................................................... 71

7-1 q-ary Compressive Sensing: Quantization/Number of measurements tradeoff. 88
7-2 Q-ary sensing for \(q = 3\). ............................................................................. 89
7-3 Image recovery with q-ary CS. (a) Original image. (b) Reconstruction with no-noise: \(q = 2^4\), SNR = 20.2 dB. (c) Reconstruction with no-noise: \(q = 2\), SNR = 16.2 dB. (d) Reconstruction with noise: \(q = 2^5\), SNR = 18.3 dB. (e) Reconstruction with noise: \(q = 2\), SNR = 15 dB. ................. 90

8-1 The locality of the random max-out feature expected kernel \(q = 2^3\). ........ 99

9-1 The level sets of the objective of problem the 1bitPhase problem (in red maximal values, in blue minimal values). .................................................. 120
9-2 \(\lambda\) as a signal to noise ratio: \(\lambda\) versus different parameters of various observation model \(\theta\). \(\lambda\) decreases as noise and distortion levels increase. .......... 122
9-3 The level sets of the objective of the initialization: \(x^* \frac{1}{m} \sum_{i=1}^{m} b_i a_i a_i^* x\), for \(x_0 = (1, 0)\). (in red maximal values, in blue minimal values). .......... 127
9-4 Robustness of One Bit Phase Retrieval to extreme distortion: the error of recovery $1 - |\langle x, x_0 \rangle|^2$ versus the distortion level $\alpha$ .................................................... 129

10-1 Phase transition comparison of one bit Phase retrieval and SubExp phase Retrieval. ................................................. 154

10-2 Phase transition comparison of one bit Phase retrieval and SubExp phase Retrieval. ................................................. 155

10-3 Phase transition comparison of one bit Phase retrieval and SubExp phase Retrieval. ................................................. 156

10-4 Alternating minimization convergence with different initializations: Random Initialization, 1bitPhase, and SubExpPhase, in the noisy and noiseless setting. 157

10-5 An image of Stata Center (Structured) and its Power Spectra (we plot the logarithm of the power spectra of one color Channel (R for instance)). . . . 158

10-6 A drawing of Stata Center (texture like image) and its Power Spectra (we plot the logarithm of the power spectra of one color Channel (R for instance)). 158

10-7 One Bit Coded Diffraction Patterns obtained by quantizing pairs of coded diffractions patterns. ........................................... 159

10-8 Stata Center's power spectra undergoing a distortion such as clipping of the values $\tanh(\alpha|\hat{x}(w)|)$. The distorted power spectra is obtained for $\alpha = 0.001$. The severely distorted is obtained for $\alpha = 0.1$. ......................... 160

10-9 One Bit Phase Retrieval from distorted CDP. (a) and (b) Recovery from small distortions. (c) Recovery from Severe Distortions. ............................. 160

10-10 One Bit Phase Retrieval from distorted CDP. (a) and (b) Recovery from small distortions. (c) Recovery from Severe Distortions. ............................. 161

10-11 Super-Resolution and Blind Deconvolution via Phase Retrieval from One Bit CDP. ................................................................. 162

10-12 Alternating Minimization and One Bit Phase Retrieval from Gaussian masks. 163

10-13 Alternating Minimization and One Bit Phase Retrieval from Bernoulli masks. 163
List of Tables

2.1 Surrogate losses in binary classification. ........................................... 24
2.2 Simplex loss functions defined by the three different notion of multiclass geometric margin. ............................................................ 28
3.1 Simplex loss functions defined by the different notions of multiclass geometric margin. ............................................................... 31
5.1 Accuracies of our algorithms on several datasets. .............................. 64
6.1 Empirical Evaluation on the AV-ASR Studio dataset. ....................... 72
6.2 Empirical evaluation on the AV-ASR Studio dataset. ....................... 76
8.1 Random Maxout on MNIST. ............................................................ 102
9.1 Comparison of the sample and computational complexity of different phase retrieval schemes. ....................................................... 118
Chapter 1

From Bits to Information

As more and more data is available, robust, memory and time efficient algorithms become critical in many data driven application fields. The design of such algorithms is central in supervised learning and modern signal acquisition, representation and reconstruction.

On the first hand, in computer vision or speech recognition, the task of classifying signals requires finding good representations of various available signal modalities, as well as defining a notion of localized similarity between them, in order to assign correctly discrete bits, referring to the semantic class of the signal of interest. In information retrieval, it is of great interest to represent the data in terms of short codes defined on a q-ary alphabet in such a way that the Hamming distance between the codes preserves a similarity metric defined on the original data. This would allow for fast search using approximate nearest neighbor algorithms.

On the other hand, in many applications in computational photography and imaging science, pushing the dynamic range of the acquired signals to the level of a bit offers a large venue of robustness to several sources of noise and distortions, as well as memory efficiency, and poses new challenges in solving those new discrete quantized inverse problems.

In this thesis we focus on those aspects of quantization in machine learning and modern signal acquisition and representation and its implication on a statistical algorithm, in terms of robustness as well as memory, sample and time complexities.

A key feature of the approach taken in this thesis is representing signals in terms of the order statistics of available non linear measurements, rather than relying on their actual values. Non linear measurements in semantic, compressed and redundant representations of signals, are the key for the success of many multi-layered, non-linear data representations in machine learning.
1.1 Problem Area and State of the Field

Motivated by the applications presented in the introduction, we study in this thesis quantized semantic, compressive and redundant representations, i.e., vector representations of signals whose entries are restricted to belong to a finite alphabet. For the purpose of illustration, we consider in this section the alphabet $\mathcal{Y} = \{0, 1, 2, 3\}$, that we encode with four colors, red, green, blue, and yellow respectively.

**Multiclass and Multimodal Classification or Semantic Representations.** Multiclass learning problems are ubiquitous in machine learning. In the unimodal supervised learning setting, a $q$-class classification problem is defined as follows: we have a training set $S$ of labeled signals $S = \{(x_i, y_i), x_i \in \mathcal{X}, y_i \in \mathcal{Y} = \{0, \ldots, q - 1\}, i = 1 \ldots N\}$, and we would like to learn a classification rule $b : \mathcal{X} \rightarrow \mathcal{Y}$ (See Figure 1-1).

While theory and algorithms for solving binary classification problems ($q = 2$) are well established [BJM06], the problem of multi-category classification is much less understood. Practical multiclass algorithms often reduce the problem to a collection of binary classification problems [RK04]. Extending statistical consistency results to the multiclass setting is not straightforward; We address this problem in this thesis by proposing a new geometric relaxation for the multiclass problem.

![Figure 1-1: A unimodal multi-class problem consists in learning a model that predicts the semantic class label of a unimodal signal.](image)

Turning now to the multimodal setting, it consists of fusing and relating information coming from different sources. Finding correlations between different modalities, and modeling their interactions, is at the heart of multimodal learning [YGS89, NKK+11, SS12, SCMN13, AALB13]. An important example of such a setting is the Audio-Visual Automatic Speech Recognition problem (AV-ASR). The multimodal signal in this case consists of the speech audio-waveforms and the lip motion video clips (see Figure 1-2). The classification consists in predicting the correct phoneme class label given this multimodal signal. We address in
Figure 1-2: An example of multimodal multi-class problems is the Audio-Visual Automatic Speech Recognition problem (AV-ASR).

this thesis this problem, in particular the modeling of the non-linear correlation between the audio and the visual channel.

Compressive Sensing or Compressed Representations. Compressive sensing of structured signals consists of measuring a lower dimensional representation of the signal [CT06]. The information content of this compressed representation allows for the recovery of the original signal and to approximately compute distances between signals. Consider a structured signal \( x_0 \in \mathbb{R}^n \), and random Gaussian matrix \( A \in \mathbb{R}^{m \times n} \), where \( m << n \), a compressed representation is obtained by measuring \( C(x_0) = Ax_0 \). The encoding \( C \) defines a new paradigm in the acquisition of structured signals known as compressive sensing [CT06], and in fast computation of approximate pairwise distances known as locality sensitive hashing [IN07, AI08].

Figure 1-3: Compressive acquisition of a scene on the alphabet \( \{0, 1, 2, 3\} \).

Quantizing random projections is a step forward in bypassing the classic pipeline of analogue acquisition of linear measurements followed by scalar quantization techniques and substituting it by a compressive acquisition of non-linear quantized measurements, such as
the so called one bit measurements: \( y = C_1(x_0) = \text{sign}(Ax_0) \) [BB08]. The encoding \( C_1 \) results in a compressive representation defined on the alphabet \( \mathcal{Y} = \{-1, 1\} \), which on one hand is widely used in approximate nearest neighbor search [IN07], and on the other hand, recovering \( x_0 \) from those binary strings is a natural question raised in the growing field of one bit compressive sensing [BB08, PV13b, PV13a, JLBB13]. In this thesis we propose and analyze non-linear compressive representation of structured signals on a general alphabet \( \mathcal{Y} = \{0, \ldots, q - 1\} \) (See Figure 1-3).

**Phase Retrieval or Redundant Representations.** In various situations we have access to measurements of a signal with missing information. For instance, the phase retrieval problem, where we have access to phase-less linear measurements i.e we measure \( C(x_0) = |Ax_0| \), where \( x_0 \in \mathbb{C}^n, A \in \mathbb{C}^{m \times n} \). In order to recover \( x_0 \), \( C \) needs to be a redundant representation of \( x_0 \), meaning that the encoding dimension \( m \) is larger than the ambient dimension \( n \). The problem of phase retrieval is ubiquitous in many areas of imaging science and engineering, where we are able to measure only the magnitude of measurements. Such a problem arises for example in X-ray crystallography [Har93, Lea08], diffraction imaging [BDP+07, Rod08] or microscopy [MISE08], where one can only measure the intensities of the incoming waves, and wishes to recover the lost phase in order to be able to reconstruct the desired object. Recovering signals from the magnitude of their Fourier transform is at the core of the phase retrieval problem.

![Figure 1-4: A redundant representation defined on a finite alphabet in a x-ray diffraction imaging.](image)

In practice, phase recovery is a non convex problem and often tackled via greedy algorithms [GS72, Fie82, GL84] which typically lack convergence guarantees. Recently, approaches based on convex relaxations, namely PhaseLift in [CSV11, DH12], and Phase cut in [WDM12], have been proposed and analyzed. These latter methods can be solved by Semi Definite Programing (SDP), and allow the exact and stable recovery of the signal (up to a global phase) from \( O(n) \) measurements. Another approach based on suitably initialized non convex optimization has been shown to be provably effective in phase retrieval in [NJS13, CLS14]. In this thesis we ask the question of robust phase recovery, when the mag-
magnitude measurements are undergoing a possibly unknown non linearity $\theta$, i.e., we measure
\[
C(x_0) = \theta(|Ax_0|),
\]
and we find that this problem can be effectively solved via a comparative encoding, based on quantizing the magnitude measurements on a finite alphabet (See Figure 1-4).

1.2 Thesis Contributions

We summarize in the following the principal thesis contributions:

**Multiclass and Multimodal Classification or Semantic Representations.** We develop in Chapters 2, 3, 4, and 5, a new theory for multiclass classification based on a coding/decoding strategy and derive algorithms within the framework of kernel methods or regularization in vector valued Reproducing Kernel Hilbert Spaces. We validate our algorithms on various benchmarks and computer vision classification tasks. On the other hand we develop in Chapter 6 a new neural network in the deep multimodal learning framework, that models the non-linear correlations between different modalities, and that predicts the semantic class of multimodal signals of interest. We validate our approach on Audio-Visual Automatic Speech Recognition (AV-ASR) on the IBM AV-ASR Large Vocabulary Studio dataset.

**Compressive Sensing or Compressed Representations.** In Chapter 7 we study an embedding of the unit sphere to a $q$-ary Hamming cube, and the corresponding recovery problem i.e recovering a unitary, structured (for e.g. sparse) signal from the order statistics of linear measurements baptized as $q$-ary Compressive Sensing. Based on the order statistics of random linear measurements we introduce in Chapter 8, the random maxout feature map that allows for locally linear estimation of functions of interest in a statistical learning task. We validate our approach as lossy compression of images, as well as a locality sensitive hashing. Finally we validate the use of random Maxout feature map in a hand written digit classification task.

**Phase Retrieval or Redundant Representations.** We introduce and study in Chapters 9 and 10, a binary quantization scheme for non negative phase-less measurements, that allows a robust recovery of the signal from the order statistics of phase-less (quadratic) measurements. We call this approach One Bit Phase Retrieval. We study and analyze our algorithms within the Gaussian sampling framework as well as in a more realistic setting, namely the Fourier sampling, of particular interest in X-ray Crystallography, and diffraction imaging. We present applications of our algorithms, and validate their robustness and efficiency in phase retrieval and super-resolution in diffraction imaging.
Part I

Multiclass & Multimodal Learning
Chapter 2
Multiclass Learning via Simplex Coding

In this chapter we introduce a geometric relaxation for the combinatorial multiclass classification problem. Our relaxation is based on a coding/decoding strategy namely the simplex coding.

2.1 Introduction

Multiclass learning is becoming increasingly important in machine learning, as bigger and bigger datasets become available. While theory and algorithms for solving binary classification problems are well established, the problem of multi-category classification is much less understood. Practical multiclass algorithms often reduce the problem to a collection of binary classification problems. Binary classification algorithms are often based on a relaxation approach: classification is posed as a non-convex minimization problem and hence relaxed to a convex one, defined by suitable convex loss functions. In this context, results in statistical learning theory quantify the error incurred by relaxation and in particular derive comparison inequalities explicitly relating the excess misclassification risk with the excess expected loss, see for example [BJM06, YRC07, RW10, Zha04b] and Chapter 3 of [SC08] for an exhaustive presentation as well as generalizations.

Generalizing the above approach and results to more than two classes is not straightforward. Several computational solutions have been proposed (among others, see [LLW04, DB95, CS01, WW99, ASS00, TJHA05]). Most proposed methods have computational complexity which is more than linear in the number of classes and simple one versus all in practice offers a good alternative both in terms of performance and speed [RK04]. Much fewer works have focused on deriving theoretical guarantees. Results in this sense have been pioneered by [Zha04a, TB05], see also [Liu07, Gue07, VWR11]. In these works the error due to relaxation is studied asymptotically and under constraints on the function class to be considered. More quantitative results in terms of comparison inequalities are given in [CS06] under similar restrictions (see also [Tar08]). Notably, the above results show that seemingly intuitive extensions of binary classification algorithms might lead to methods which are not consistent.
Further, it is interesting to note that these restrictions on the function class, needed to prove
the theoretical guarantees, make the computations in the corresponding algorithms more
involved and are in fact often ignored in practice.
In this thesis we discuss a novel framework for multiclass learning, defined by a suitable cod-
ing/decoding strategy, namely the simplex coding, in which a relaxation error analysis can
be developed avoiding constraints on the considered hypotheses class. Moreover, we show
that in this framework it is possible to derive consistent regularized method with tuning
complexity which is independent to the number of classes.

2.2 Background

Let \((X, Y)\) be two random variables with values in a measurable space \(\mathcal{X}\) and \(\mathcal{Y} = \{0 \ldots q - 1\}\), respectively, where \(q \in \mathbb{N}, q \geq 2\). We will denote by \(\rho_X\), the law of \(X\) on \(\mathcal{X}\), and by \(\rho_j(x)\), the conditional probabilities \(P(Y = j | X = x)\) for \(j \in \mathcal{Y}\). We can think of \(\mathcal{X}\) as a set
of possible inputs and of \(\mathcal{Y}\), as a set of labels describing a set of semantic categories/classes
the input can belong to.
A classification rule is a map \(b : \mathcal{X} \rightarrow \mathcal{Y}\), and its error is measured by the misclassification
risk:

\[ R(b) = \mathbb{P}(b(X) \neq Y) = \int_{\mathcal{X}} \sum_{j=0}^{q-1} 1_{b(x) \neq j} \rho_j(x) d\rho_X(x). \]  \hspace{1cm} (2.1)

We would like to find the classification rule that minimizes the misclassification risk:

\[ \min_b R(b). \]  \hspace{1cm} (2.2)

The optimal classification rule that minimizes \(R\) is the Bayes rule,

\[ b_\rho(x) = \arg \max_{y \in \mathcal{Y}} \rho_y(x), \ x \in \mathcal{X}. \]  \hspace{1cm} (2.3)

Computing the Bayes rule by directly minimizing (2.1), is not possible since the probability
distribution is unknown.
In practice we have access to a training set \(S\) consisting of pairs of inputs \(x_i\) and labels \(y_i,\)
\(S = \{(x_i, y_i), i = 1 \ldots N\}\). We can therefore write the following empirical problem:

\[ \hat{R}(b) = \frac{1}{N} \sum_{i=1}^{N} 1_{b(x_i) \neq y_i}. \]  \hspace{1cm} (2.4)

The problem in (2.4) corresponds to a discrete and non differentiable optimization problem.
In binary classification, one of the most common way to obtain computationally efficient
methods is based on a relaxation approach. We recall this approach in the next section and
describe in this chapter how to extend it to the multiclass setting.
2.2.1 Relaxation Approach to Binary Classification

In binary classification we have $q = 2$ and we can set $\mathcal{Y} = \pm 1$. Most modern machine learning algorithms for binary classification consider a convex relaxation of the functional in Eq. (2.1). More precisely:

1. The indicator function in Eq. (2.1) is replaced by a non negative loss $V : \mathcal{Y} \times \mathbb{R} \to \mathbb{R}^+$ which is convex in the second argument and is sometimes called a surrogate loss.

2. The classification rule $b$ is replaced by a real valued measurable function $f : \mathcal{X} \to \mathbb{R}$. A classification rule is then obtained considering the sign of $f$, i.e. for each $x \in \mathcal{X}$ we have $b(x) = \text{sign}(f(x)) = 1$ if $f(x) > 0$ and $-1$ otherwise.

In fact, typically it suffices to consider a special class of loss functions, namely large margin loss functions $V : \mathbb{R} \to \mathbb{R}^+$ of the form $V(yf(x))$. This last expression is suggested by the observation that the misclassification risk, using the labels $\pm 1$, can be written as,

$$R(f) = \mathbb{E}(\Theta(-Yf(X))),$$

where $\Theta$ is the heavy side step function, i.e. $\Theta(x) = 1$, if $x > 0$ and $0$ otherwise. The quantity $m = yf(x)$, sometimes called the margin, is a natural pointwise measure of the classification error.

We recall in the following table popular surrogate loss functions, and their corresponding algorithms in the binary classification literature (see also Figure 2-1):
2.2.2 Relaxation Error Analysis for Binary Classification

As we replace the misclassification loss with a convex surrogate loss $V$, we are effectively changing the problem, since the misclassification risk is replaced by the expected loss, or the expected $V$-Risk,

$$ \mathcal{E}_V(f) = \mathbb{E}(V(Yf(X))), $$

induced by $V$.

The expected loss depends on $V$ and the data distribution $\rho$, let $f_\rho$ be the minimizer of $\mathcal{E}_V$:

$$ f_\rho = \arg \min_f \mathcal{E}_V(f). \quad (2.5) $$

In this section we recall interesting results from [BJM06], that answer the following question: What is the relationship between $f_\rho$ and $b_\rho$? What is the error we incur by replacing the misclassification risk $R$, by the expected $V$-risk $\mathcal{E}_V$?

The relaxation error for a given loss function can be quantified by the following two requirements:

- **Fisher Consistency/classification calibration.** We say that a loss is Fisher consistent or classification calibrated, if the corresponding target function satisfies $\text{sign}(f_\rho(x)) = b_\rho(x)$ almost surely.

- **Comparison inequalities/ Bayes Consistency.** The excess missclassification risk, and the excess expected risk are related by a comparison inequality

$$ R(\text{sign}(f)) - R(b_\rho) \leq \psi(\mathcal{E}(f) - \mathcal{E}(f_\rho)), $$

for any arbitrary function $f$, where $\psi = \psi_{V,\rho}$ is a suitable function that depends on $V$, and possibly on the data distribution. In particular $\psi$ should be such that $\psi(s) \to 0$ as $s \to 0$, so that if $\tilde{f}_N$ is a (possibly random) sequence of functions, such that $\mathcal{E}(\tilde{f}_N) \to \mathcal{E}(f_\rho)$ (possibly in probability), then the corresponding sequences of classification rules $\tilde{b}_N = \text{sign}(\tilde{f}_N)$ is Bayes consistent, i.e $R(\tilde{b}_N) \to R(b_\rho)$ (possibly in probability). More, generally $\psi$ should be such that a finite sample bound on the excess expected risk immediately yields one on the excess missclassification risk.

<table>
<thead>
<tr>
<th>Loss</th>
<th>$V(yf(x))$</th>
<th>Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hinge Loss</td>
<td>$\max(1 - yf(x), 0)$</td>
<td>Support Vector Machines [Vap98]</td>
</tr>
<tr>
<td>Logistic Loss</td>
<td>$\log(1 + e^{-yf(x)})$</td>
<td>Logistic Regression, GentleBoost [FS96]</td>
</tr>
<tr>
<td>Exponential Loss</td>
<td>$e^{-yf(x)}$</td>
<td>Adaboost [FS96]</td>
</tr>
<tr>
<td>Least Squares</td>
<td>$(y - f(x))^2 = (1 - yf(x))^2$</td>
<td>Regularized Least Squares [Wah90]</td>
</tr>
</tbody>
</table>

Table 2.1: Surrogate losses in binary classification.
The relaxation error in the binary case has been thoroughly studied in \cite{BJM06, RW10}. In particular, Theorem 2 in \cite{BJM06} shows that if a large margin surrogate loss is convex, differentiable and decreasing in a neighborhood of 0 ($V'(0) < 0$), then the loss is Fisher consistent. Moreover, in this case it is possible to give an explicit expression of the function $\psi$.

In particular, for the hinge loss the target function $f_\rho$ is exactly the Bayes rule $b_\rho$ and $\psi(t) = |t|$. For least square the target function $f_\rho(x) = 2\rho_1(x) - 1$, and $\psi(t) = \sqrt{t}$. The comparison inequality for the square loss can be improved for a suitable class of probability distribution satisfying the so called Tsybakov noise condition \cite{Tsy04},

$$\mathbb{P}(\{x \in \mathcal{X}, |f_\rho(x)| \leq s\}) \leq B_\beta s^\beta, \ s \in [0, 1], \beta > 0, \ B_\beta > 0. \quad (2.6)$$

Under this condition the probability of points mapped near 0 decreases polynomially. In this case the comparison inequality for the square loss is given by $\psi(t) = c_\beta t^{\frac{\beta + 1}{\beta}}$, see \cite{BJM06, YRC07}.

### 2.2.3 Previous Work on Multiclass Classification

From a practical perspective, several computational solutions to multiclass learning have been proposed. Among others, we mention for example \cite{LLW04, DB95, CS01, WW99, ASS00, TJHA05}. Indeed, most of the above methods can be interpreted as a relaxation of the original multiclass problem.

Interestingly, the study in \cite{RK04} suggests that the simple one-vs-all scheme should be a practical benchmark for multiclass algorithms as it seems to experimentally achieve performances that are similar or better than more sophisticated methods.

As we previously mentioned from a theoretical perspective a general account of a large class of multiclass methods has been given in \cite{TB05}, building on results in \cite{BJM06} and \cite{Zha04a}. Notably, these results show that seemingly intuitive extensions of binary classification algorithms might lead to inconsistent methods.

In \cite{Liu07, VWR11}, authors developed consistent multiclass SVM. When we consider $q$ classes, the classification rule is found by applying a suitable prediction/decoding map to a function

$$f : \mathcal{X} \rightarrow \mathbb{R}^q$$

where $f$ is found considering the minimization of a loss function $V : \mathcal{Y} \times \mathbb{R}^q \rightarrow \mathbb{R}^+$. The considered functions have to satisfy the so called sum to zero constraint

$$\sum_{\nu \in \mathcal{Y}} f^\nu(x) = 0, \text{ for all } x \in \mathcal{X}. \quad (2.7)$$
The latter requirement is problematic since it makes the computations in the corresponding algorithms more involved and is in fact often ignored, so that practical algorithms often come with no consistency guarantees. In all the above papers relaxation is studied in terms of Fisher and Bayes consistency and the explicit form of the function $\psi$ is not given. More quantitative results in terms of explicit comparison inequality are given in [CS06] and (see also [Tar08]), but also need to impose the sum to zero constraint on the considered function class.

2.3 A Relaxation Approach to Multi-Class Classification

2.3.1 Simplex Coding

In this section we propose a natural extension of the relaxation approach that avoids constraining the class of functions to be considered, and allows to derive explicit comparison inequalities.

We start by the first ingredient that is the coding/decoding strategy. This strategy is described by the following definition.

**Definition 1 (Simplex Coding/Decoding).** The simplex coding map is $C : \mathcal{Y} \to \mathbb{R}^{q-1}$, $C(y) = c_y$, where $C = \{c_y : y \in \mathcal{Y}\}$ satisfy:

1) $\|c_y\|^2 = 1$,
2) $\forall y, y' \in \mathcal{Y}$ $\langle c_y, c_{y'} \rangle = -\frac{1}{q-1}$,
3) $\sum_{y \in \mathcal{Y}} c_y = 0$.

The corresponding decoding is the map $D : \mathbb{R}^{q-1} \to \mathcal{Y}$, $D(\alpha) = \arg \max_{y \in \mathcal{Y}} \langle \alpha, c_y \rangle$, $\forall \alpha \in \mathbb{R}^{q-1}$.

![Decoding with simplex coding $q = 3$.](image)

The simplex coding corresponds to the $q$ most separated vectors on the hypersphere $S^{q-2}$ in $\mathbb{R}^{q-1}$, that is the vertices of the simplex (see Figure 2-2). For binary classification it reduces to the $\pm 1$ coding and the decoding map is equivalent to taking the sign of $f$. 

26
2.3.2 A Geometric Relaxation approach

The following steps are the key contributions of our geometric relaxation:

1. **Code**: We replace the labels by their codes i.e. \( C(y) = c_y \).

2. **Lift**: Instead of looking for a discrete classification rule \( b \) we look for a vector valued function \( f \) taking values in \( \mathbb{R}^{q-1} \), where we can go from one representation to another, with the coding/decoding maps \((C, D)\).

3. **Embed**: We embed the classification problem to a geometric problem, that consists in mapping the input space \( \mathcal{X} \), to an output space of dimension \( q - 1 \), with a constraint on the map \( f \) that a pair \((x, y) \in \mathcal{X} \times \mathcal{Y}\), is mapped to a neighborhood \( T_y \) of the vertex \( c_y \).

The choice of the neighborhood \( T_y, y \in \mathcal{Y} \) defines the notion of margin and determines how we account for errors in multiclass learning.

2.3.3 Multiclass Margin

In this section we provide several constructions of the neighborhoods \( T_y [\text{MPR14, MPRJ12}] \) and the corresponding margins:

1. **Half Spaces margin (HS margin)**: \( T_y = \{ w \in \mathbb{R}^{q-1} \mid \langle c_y, w \rangle \geq 1 \} \) i.e the half space delimited by the hyperplane orthogonal to \( c_y \) is associated to each label \( y \in \mathcal{Y} \).

2. **Discriminative Conic margin (DC margin)**: \( T_y = \{ w \in \mathbb{R}^{q-1} \mid \langle c_y, w \rangle \leq -\frac{1}{q - 1}, \forall y' \in \mathcal{Y}/\{y\} \} \), i.e a cone supported on the vertex \( c_y \), is associated to each label \( y \in \mathcal{Y} \).

3. **All pairs Conic margin (AC margin)**: The set \( T_y = \{ w \in \mathbb{R}^{q-1} \mid \langle c_y - c_{y'}, w \rangle \geq \frac{q}{q - 1}, \forall y' \in \mathcal{Y}/\{y\} \} \), i.e the voronoi region of the vertex \( c_y \) in \( C \) is associated to each label \( y \).

4. **Spheric margin (SLS margin)**: \( T_y = \{ w \in \mathbb{R}^{q-1}, ||c_y - w||^2 \leq 1 \} \), that corresponds to a simple least squares fitting (the radius one is arbitrary).

2.3.4 Multiclass Simplex Loss Functions

In the previous section we introduced several notions of margin corresponding to different neighborhood \( \{T_y\}_{y \in \mathcal{Y}} \). We are now ready for our second ingredient namely defining simplex multiclass loss functions. Given one of the above margin definition, a simplex loss function penalizes a misclassification whenever a pair \((x, y) \) is mapped outside the neighborhood \( T_y \).

In order to penalize errors, we need to impose on our losses that:

For all \( y, V(y, w) \leq V(y, w'), \) for \( w \in T_y \), and \( w' \notin T_y \).  \hspace{1cm} (2.8)
Figure 2-3: Different geometric margin definitions. While in HS points in the same class are mapped to a half space, in LS they are mapped to a sphere. In DC they are mapped to a cone, and in AC to a voronoi cell of the corresponding vertex.

Intuitively this condition ensures that the minimum penalty is for points belonging to the correct neighborhood.

Hence the following losses definition $V : \mathcal{Y} \times \mathbb{R}^{q-1} \rightarrow \mathbb{R}^+$, given in the following table,

<table>
<thead>
<tr>
<th>Margin</th>
<th>$V(y, w)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>HS</td>
<td>$\phi(1 - \langle c_y, w \rangle)$</td>
</tr>
<tr>
<td>DC</td>
<td>$\sum_{y' \neq y} \phi(\frac{1}{q-1} + \langle c_{y'}, w \rangle)$</td>
</tr>
<tr>
<td>AC</td>
<td>$\sum_{y' \neq y} \phi(\frac{q}{q-1} - \langle c_y - c_{y'}, w \rangle)$</td>
</tr>
<tr>
<td>SLS</td>
<td>$</td>
</tr>
</tbody>
</table>

Table 2.2: Simplex loss functions defined by the three different notion of multiclass geometric margin.
where \( \phi : \mathbb{R} \to \mathbb{R}^+ \) is a scalar function. Choices of \( \phi \), correspond for example to \( \phi(z) = \max(z, 0) \) related to the hinge loss and the popular support vector machine [Vap98], \( \phi(z) = e^z \) or the exponential loss popularized in the Boosting literature, and \( \phi(z) = \log(1 + e^z) \), the logistic loss [SY12].

Note that in all that cases condition (2.8) is satisfied. To illustrate that we show in the following the level sets of different losses under different margins in a three classes scenario.

Figure 2-4: Level sets of different simplex losses: Discriminative Conic Margin.

Figure 2-5: Level sets of different losses: All pairs Conic Margin.
2.3.5 Expected V-Risk Minimization

Having embedded the multiclass problem to this geometry we shall replace the discrete function $b$ with the vector valued function $f$, substitute the indicator function in (2.1) with a simplex loss function, and minimize the expected $V-$ loss $\mathcal{E}_V(f)$, instead of the misclassification error:

$$\min f \mathcal{E}_V(f) := \min f \mathbb{E}_{X,Y} V(Y, f(X)) = \min f \int_X \sum_{j=0}^{q-1} V(y, f(x)) \rho_j(x) d\rho_X(x). \quad (2.9)$$

Let $f_\rho$ be the minimizer of $\mathcal{E}_V$, to get a discrete classifier we can use the decoder $D$. $D(f_\rho)$ is therefore the classification rule that this method produces.

A loss function is said to be classification calibrated or fisher consistent if $D(f_\rho) = b_\rho$ (almost surely), in other words nothing is lost in expectation, the bayes rule is still the optimum of the relaxed problem.

In the next chapters we study conditions on the simplex loss functions that ensure Fisher consistency.

Indeed $f_\rho$ is a theoretical object, since we have access to only few samples of the distributions $\rho$ it is therefore important to quantify the excess $V$- risk of an arbitrary function $f$, $\Delta \mathcal{E}_V(f) = \mathcal{E}_V(f) - \mathcal{E}_V(f_\rho)$ and to relate it the excess misclassification risk induced by the classification rule $D(f)$ namely $\Delta R(D(f)) = R(D(f)) - R(D(f_\rho))$. If the loss was fisher consistent the latter is equal to $R(D(f)) - R(b_\rho)$. In the following chapters we prove for particular loss functions that we have explicit comparison inequalities having the following form:

$$R(D(f)) - R(b_\rho) \leq \psi (\mathcal{E}_V(f) - \mathcal{E}_V(f_\rho)), \quad (2.10)$$

where $\psi$ is an increasing scalar function. Hence, we obtain Bayes consistency, i.e if we have a sequence of function $\hat{f}_N$, such that $\mathcal{E}_V(\hat{f}_N) \to \mathcal{E}_V(f_\rho)$, then $R(D(\hat{f}_N)) \to R(b_\rho)$

2.4 Conclusion

In this chapter we introduced a geometric relaxation for multiclass learning via a coding/decoding strategy using the simplex coding. We defined several notions of margin in order to account for misclassification errors, and introduced simplex loss functions $V$ that reflect that geometric margin. Following the convex relaxation [BJM06] in binary classification we can now minimize the expected $V-$ risk, instead of the misclassification error.

In the next chapter we answer the following question:

*What is the price of such relaxation in term of Fisher consistency?*
Chapter 3

Consistency Analysis of Simplex Multiclass Learning

In this chapter we study the consistency of the geometric relaxation, based on the simplex coding introduced in Chapter 2.

3.1 Introduction

As discussed in Chapter 2, given the margin definitions, we can construct several surrogate losses $V$ for the multiclass problem. In Definition (2), we give a precise definition of a multiclass simplex loss function:

**Definition 2** (Multiclass Simplex Loss Function). A multiclass simplex loss function $V$ is a function $V : \mathcal{Y} \times \mathbb{R}^{q-1} \rightarrow \mathbb{R}^+$, that has one of the following forms:

$$V(y, w) = \phi(1 - \langle c_y, w \rangle)$$

$\phi : \mathbb{R} \rightarrow \mathbb{R}^+$ is a scalar (convex) function.

Choices of $\phi$, correspond for example to $\phi(z) = \max(z, 0) := |z|_+$, related to the hinge loss and the popular Support Vector Machines (SVM) [Vap98], $\phi(z) = e^z$ or the exponential loss popularized in the Boosting literature, and $\phi(z) = \log(1 + e^z)$, the logistic loss [SY12].

Following the relaxation approach discussed in Chapter 2, we shall replace the misclassification risk $R$, by the risk induced by the loss $V$:

Table 3.1: Simplex loss functions defined by the different notions of multiclass geometric margin.

<table>
<thead>
<tr>
<th>Margin</th>
<th>$V(y, w)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>HS</td>
<td>$\phi(1 - \langle c_y, w \rangle)$</td>
</tr>
<tr>
<td>DC</td>
<td>$\sum_{y' \neq y} \phi\left(\frac{1}{q-1} + \langle c_{y'}, w \rangle\right)$</td>
</tr>
<tr>
<td>AC</td>
<td>$\sum_{y' \neq y} \phi\left(\frac{q}{q-1} - \langle c_y - c_{y'}, w \rangle\right)$</td>
</tr>
</tbody>
</table>
Definition 3 (Expected V-Risk). Given a multiclass simplex loss function $V$, let $f$ be a measurable vector valued function $f : \mathcal{X} \rightarrow \mathbb{R}^{q-1}$, we define the expected V-Risk as:

$$E_V(f) = \mathbb{E}(V(Y, f(X))) = \int_{\mathcal{X}} \sum_{j=0}^{q-1} \rho_j(x)V(j, f(x))d\rho_X(x). \quad (3.1)$$

It will be useful in our analysis to introduce the following definition:

Definition 4 (Inner V-Risk). Let $\Delta_q = \left\{ p \in [0,1]^q \mid \sum_{j=0}^{q-1} p_j = 1 \right\}$ be the probability simplex. Let $V$ be a multiclass simplex loss function, $p \in \Delta_q$ a probability vector, and $w \in \mathbb{R}^{q-1}$, we define the inner V-risk:

$$\Psi_V(w, p) = \sum_{j=0}^{q-1} p_j V(j, w). \quad (3.2)$$

and let $H(p)$ be its infimum:

$$H(p) = \inf_{w \in \mathbb{R}^{q-1}} \Psi_V(w, p). \quad (3.3)$$

Note that we have, for any measurable function $f : \mathcal{X} \rightarrow \mathbb{R}^{q-1}$:

$$E_V(f) = \int_{\mathcal{X}} \Psi_V(f(x), p(x))d\rho_X(x) = \mathbb{E}(\Psi_V(f(X), P(X))), \quad (3.4)$$

where $p(x) = (\rho_0(x), \ldots, \rho_{q-1}(x)), x \in \mathcal{X}$.

The optimal V-Risk is given by:

$$E^*_V = \inf_f E_V(f) = \inf_f \mathbb{E}(\Psi_V(f(X), P(X))), \quad (3.5)$$

where the infimum is taken over all measurable functions.

A sequence $w_1, w_2, \ldots$ achieves $H(p)$ if:

$$\lim_{i \to \infty} \Psi_V(w_i, p) = H(p). \quad (3.6)$$

If the infimum is achieved by a unique $w^*$, we can in that case define $w^* : \Delta_q \rightarrow \mathbb{R}^{q-1}$, such that:

$$w^*(p) = \arg \min_{w \in \mathbb{R}^{q-1}} \Psi_V(w, p). \quad (3.7)$$

In that case we can define $f^*_V$ the minimizer of $E_V, f^*_V : \mathcal{X} \rightarrow \mathbb{R}^{q-1}$ (up to $\rho_X$- null sets):

$$f^*_V(x) = \arg \min_{w \in \mathbb{R}^{q-1}} \Psi_V(w, p(x)). \quad (3.8)$$

As discussed in Chapter 2, the minimizer of the misclassification risk $R$ is the bayes rule:

$$b_p(x) = \arg \max_{j=0\ldots q-1} \rho_j(x) \quad x \in \mathcal{X}.$$
For the arg max ties are broken arbitrarily (otherwise equalities are replaced with inclusions). A classification rule should be consistent with the Bayes rule. For \( p \in \Delta_q \), define \( H^- \):

\[
H^-(p) = \inf_{w, D(w) \not\in \arg\max_{j \in \mathcal{Y}} p_j} \Psi_V(w, p). \tag{3.9}
\]

\( H^-(p) \) is the infimum of \( \Psi_V(., p) \), under the constraint that the classification rule induced by \( w (D(w)) \) is inconsistent with the Bayes rule. We are now ready to define the requirement on \( V \), in order to have consistency with the Bayes rule:

**Definition 5** (Classification Calibration). A multiclass simplex loss function \( V \) is classification calibrated if for any \( p \), such that \( p_i \neq p_j \) for all \( i, j \in \mathcal{Y}, i \neq j \):

\[
H^-(p) > H(p). \tag{3.10}
\]

If \( w_1, w_2, \ldots \) was a sequence achieving \( H(p) \), this means that this sequence satisfies:

\[
D(w_i) \neq \arg\max_j p_j, \tag{3.11}
\]

only finitely often.

Equivalently, if the infimum of \( H \) was achieved by a unique \( w^* \), or equivalently the infimum of \( \mathcal{E}_V \) was uniquely achieved by \( f^*_V \). Then this condition means that \( f^*_V \) satisfies:

\[
D(f^*_V) = b_p, \tag{3.12}
\]

almost surely (this was the definition of Fisher consistency in Chapter 2).

### 3.2 Classification Calibration

We start with preliminary definitions on the confusion zones of the simplex code, this will be essential in characterizing the classification calibration of a simplex loss function.

**Definition 6** (Confusion Zones). For \( y, y' \in \mathcal{Y}, y \neq y' \), we define a confusion zone of the simplex code:

\[
R_{y, y'} = \{ w \in \mathbb{R}^{q-1} | \langle c_y, w \rangle = \langle c_{y'}, w \rangle \}. \tag{3.13}
\]

We note by \( \Delta^y_q \) the set of (distincts) probabilities \( p = (p_0 \ldots p_{q-1}) \) having \( p_y \) as their maximum:

\[
\Delta^y_q = \left\{ p \in \Delta_q | y = \arg\max_{j \in \mathcal{Y}} p_j \right\}, \tag{3.14}
\]

and the Voronoi cell of a vertex \( c_y \):

\[
S_y = \{ w \in \mathbb{R}^{q-1} | \langle c_y, w \rangle > \langle c_{y'}, w \rangle, \forall y' \in \mathcal{Y}, y' \neq y \}. \tag{3.15}
\]
For a point $w \in S_y$, we define the following map that maps $w$, to a point in the confusion zone $R_{y,y'}$:

$$w' = \Pi_{y,y'}(w) = \frac{q-1}{q} \left( \frac{(c_y + c_{y'}, w)}{2} (c_y + c_{y'}) + \sum_{j \neq y, y'} (c_j, w) c_j \right). \quad (3.16)$$

It is easy to verify that $w'$ satisfies:

$$\langle c_y, w' \rangle = \langle c_{y'}, w' \rangle = \left\langle \frac{c_y + c_{y'}}{2}, w \right\rangle, \text{ and } \langle c_j, w' \rangle = \langle c_j, w \rangle, j \in \mathcal{Y}, j \notin \{y, y'\}. \quad (3.17)$$

For $p \in \Delta_q$, and $y, y' \in \mathcal{Y}$ and $y' \neq y$, such that $p_y \neq p_{y'}$. We define the following useful probabilities

1. $p' \in \Delta_q$ such that $p'_y = p_{y'} = \frac{p_y + p_{y'}}{2}$, and $p'_k = p_k, k \in \mathcal{Y}, k \notin \{y, y'\}. \quad (3.18)$

2. $p^\circ \in \Delta_q$ such that $p_{y'} = p_{y'}$, and $p_{y'}^\circ = p_y$, and $p_k^\circ = p_k, k \in \mathcal{Y}, k \notin \{y, y'\}. \quad (3.19)$

We start with the following preliminary Lemma, that relates classification calibration to a bound between the optimal V-risk evaluated for a probability vector $p$, and the optimum evaluated at $p'$. $p'$ highlights the confusion between classes at the level of the probability simplex.

**Lemma 1 (Calibration, Risk, Probability simplex).** Let $p \in \Delta_q$, and $y, y' \in \mathcal{Y}$, $y \neq y'$, such that $p_y \neq p_{y'}$. Let $p'$ be defined as in (3.18). $V$ is classification calibrated then $H(p) < H(p')$, for all $p \in \Delta_q$.

We note by $\partial \Psi_V(w, p)$, the subgradient of $\Psi_V(., p)$ at $w$. For $w \in \mathbb{R}^{q-1}$, let $w' = \Pi_{y,y'}(w)$, in the following lemma we give an important quantity $\langle \partial \Psi_V(w', p), w - w' \rangle$ that depicts the behavior of the inner loss in the confusion zone of the simplex code. This quantity is studied for various margins definitions.

**Lemma 2 (Inner Risk and Confusion Zones).** Let $\phi'$ be the sub-differential of $\phi$ (not necessarily unique at a given point). For any $y, y' \in \mathcal{Y}$, $y \neq y'$, $w \in \mathbb{R}^{q-1}$, and $w' = \Pi_{y,y'}(w)$ we have:

1. HS Margin. Let $\lambda_y(w') = \frac{1}{2} \phi'(1 - \langle c_y, w' \rangle)$, and $\lambda_{y'}(w') = \frac{1}{2} \phi'(1 - \langle c_{y'}, w' \rangle)$, then:

$$\langle \partial \Psi_V(w', p), w - w' \rangle = (-p_y \lambda_y(w') + p_{y'} \lambda_{y'}(w')) \langle c_y - c_{y'}, w \rangle.$$

For differentiable $\phi'$ we have:

$$\langle \partial \Psi_V(w', p), w - w' \rangle = \lambda_y(w') (p_{y'} - p_y) \langle c_y - c_{y'}, w \rangle.$$
2. **DC Margin.** Let \( \lambda_y(w') = \frac{1}{2} \phi'\left( \frac{1}{q-1} + (c_y, w') \right) \), and \( \lambda_{y'}(w') = \frac{1}{2} \phi'\left( \frac{1}{q-1} + (c_{y'}, w') \right) \), then:

\[
\langle \partial \Psi_V(w', p), w - w' \rangle = ((1 - p_y)\lambda_y(w') - (1 - p_{y'})\lambda_{y'}(w')) \langle c_y - c_{y'}, w \rangle.
\]

For differentiable \( \phi \) we have:

\[
\langle \partial \Psi_V(w', p), w - w' \rangle = \lambda_y(w')(p_y - p_y') \langle c_y - c_{y'}, w \rangle.
\]

3. **AC Margin.** For all \( w \in \mathbb{R}^{q-1} \), there exists \( \lambda_j(w') \), \( j \in \mathcal{Y} \), such that:

\[
\langle \partial \Psi_V(w', p), w - w' \rangle = \langle c_y - c_{y'}, w \rangle \left( -\lambda_y(w')p_y + \lambda_{y'}(w')p_{y'} + \sum_{j \neq y, y'} \lambda_j(w')p_j \right).
\]

For differentiable \( \phi \) we have:

\[
\langle \partial \Psi_V(w', p), w - w' \rangle = \lambda_y(w')(p_y - p_y') \langle c_y - c_{y'}, w \rangle.
\]

Proofs of Lemmas 1 and 2 are given in Section 3.4. We are now ready to state our main result on a condition on the simplex loss function, to have classification calibration:

**Theorem 1** (Classification Calibration). A convex loss \( V \) of type HS, DC, or AC, given in Definition 2, such that for all \( y \in \mathcal{Y} \), \( V(y, \cdot) \) is differentiable in the confusion zones \( R_{y,y'} \) for all \( y' \in \mathcal{Y}, y' \neq y \), and that satisfies for all \( y \in \mathcal{Y} \)

\[
\lambda_y(w') > 0, \quad \forall \; w' \in R_{y,y'}, \; y' \in \mathcal{Y}/\{y\}.
\]

( where \( \lambda_y(\cdot) \) are defined in Lemma 2), is classification calibrated.

**Proof of Theorem 1.** Suppose that \( V \) is convex, differentiable in the confusion zones and satisfies the condition (3.20), and that it is not classification calibration. Suppose that \( p \in \Delta_y \), \( V \) is not classification calibration implies that there exists \( w_0 \in S_{y', y' \neq y} \) that achieves the minimum. By convexity, we have:

\[
\Psi_V(w_0, p) \geq \Psi_V(\Pi_{y', y'}(w_0), p) + \langle \partial \Psi_V(\Pi_{y', y'}(w_0), p), w - \Pi_{y', y'}(w_0) \rangle
\]

\[
= \Psi_V(\Pi_{y', y'}(w_0), p) - \lambda_y(w')(p_y - p_{y'}) \langle c_y - c_{y'}, w_0 \rangle
\]

\[
> \Psi_V(\Pi_{y', y'}(w_0), p),
\]

where the equality follows from differentiability of the loss in the confusion zones and from Lemma 2. The positivity of \( -\lambda_y(w')(p_y - p_{y'}) \langle c_y - c_{y'}, w_0 \rangle \) follows from the assumptions on \( \lambda_y(\cdot), p_y, \) and \( w_0 \). \( \Psi_V(w_0, p) > \Psi_V(\Pi_{y', y'}(w_0), p) \), is a contradiction with \( w_0 \) being a minimizer. \( \square \)
3.2.1 Examples

Exponential Loss. It is easy to see that the exponential loss satisfies all conditions of Theorem 1 for all margin definitions.

Hinge Loss. In that case $\phi(z) = \max(z, 0) := |z|_+$. For DC margin, let us compute the differential of the loss, $V(y, w) = \sum_{y' \neq y} \left| \frac{1}{q-1} + \langle c_{y'}, w \rangle \right|_+$. For all $y' \in \mathcal{Y}$,

$$
\partial \left| \frac{1}{q-1} + \langle c_{y'}, w \rangle \right|_+ = \begin{cases} 
    c_{y'} & \text{if } \langle c_{y'}, w \rangle > -\frac{1}{q-1} \\
    \nu_{y'} c_{y'}, 0 < \nu_{y'} < 1, & \text{if } \langle w, c_{y'} \rangle = -\frac{1}{q-1} \\
    0, & \text{if } \langle w, c_{y'} \rangle < -\frac{1}{q-1}.
\end{cases}
$$

For $w \in R_{y,y'}$, we have always $\langle c_y, w \rangle > -\frac{1}{q-1}$, hence the loss is convex, differentiable on the confusion zone, and $\lambda_y = \frac{1}{2} \phi'\left(\frac{1}{q-1} + \langle c_{y'}, w \rangle \right) = \frac{1}{2} > 0$. Hence DC Hinge loss is classification calibrated.

3.3 Conclusion

In this Chapter we showed a condition on the loss, involving the confusion zone of the simplex code, that ensure the consistency of the geometric relaxation introduced for the multiclass problem. For now we have restricted our attention to the three margin definitions (HS, DC, AC), and we have seen the hinge loss, at the core of the so called support vector machine is consistent in the DC case, we shall call this from now on DC-SVM. The following Chapter answers the following question:

What is the relaxation error for DC-SVM?

3.4 Proofs of Lemmas 1, and 2

Proof of Lemma 1. Let $p \in \Delta_q$, and $y, y' \in \mathcal{Y}, y \neq y'$, such that $p_y \neq p_{y'}$. Let $p'$ be defined as in (3.18). Let us start by showing that $H(p) \leq H(p')$, and then show that $V$ is classification calibrated implies that $H(p) \neq H(p')$. Let $p^0 \in \Delta_q$, such that $p^0_y = p_{y'}$, and $p^0_{y'} = p_y$, and $p^0_k = p_k$, note that $p$ is a linear function in $p$, hence concave in $p$. Note that $p' = \frac{1}{2}(p + p^0)$, it follows that: $H(p') = H(\frac{1}{2} p + \frac{1}{2} p^0) \geq \frac{1}{2} H(p) + \frac{1}{2} H(p^0)$. Since our losses are symmetric around the origin:

$$
H(p) = H(p^0).
$$

It follows that:

$$
H(p') \geq H(p).
$$
Now let us prove \( V \) is classification calibrated implies that \( H(p') \neq H(p) \). The proof is a contradiction. Suppose there exists \( p \) such that \( H(p) = H(p') \). Let \( w_i \) be a sequence that achieves \( H \), at \( p' \) by definition:

\[
H(p) = \lim_{i \to \infty} \Psi_V(w_i, p) \geq H(p') = \lim_{i \to \infty} \Psi_V(w_i, p'),
\]

Hence:

\[
\left( \frac{p_y - p_{y'}}{2} \right) \lim_{i \to \infty} (V(y, w_i) - V(y', w_i)) \geq 0.
\]

Let \( p^o \in \Delta_q \), such that \( p^o_y = p_y \), and \( p^o_{y'} = p_{y'} \), and \( p^o_k = p_k \), for all \( k \neq \{y, y'\} \), since our losses are symmetric around the origin, we have:

\[
H(p) = H(p^o).
\]

The same argument shows that \( H(p) = H(p') \) implies:

\[
\left( \frac{p_y - p_{y'}}{2} \right) \lim_{i \to \infty} (V(y', w_i) - V(y, w_i)) \geq 0.
\]

It follows that:

\[
\lim_{i \to \infty} V(y, w_i) - V(y', w_i) = 0.
\]

We have finally:

\[
\lim_{i \to \infty} \Psi_V(w_i, p) = H(p),
\]

and therefore \( w_i \) achieves also \( H \) at \( p \). If \( V \) is calibrated it follows that:

\[
\lim_{i \to \infty} \text{sign} \left( (p_y - p_{y'}) (c_y - c_{y'}, w_i) \right) = 1
\]

The same argument shows that \( H \) achieved at \( p^o \) by the sequence \( w_i \), and if \( V \) is classification calibration we must have:

\[
\limsup_{i \to \infty} \text{sign} \left( (p_y - p_{y'}) (c_y - c_{y'}, w_i) \right) = -1
\]

Thus if \( H(p) = H(p') \), \( V \) is not classification calibrated. \( \square \)

**Proof of Lemma 2.** (HS Margin) We have \( V(j, w) = \phi(-\langle c_j, w \rangle) \) \( \partial V(j, w) = -\phi'(-\langle c_j, w \rangle)c_j \). Let \( w' = \Pi_{y,y'}(w) \) we shall compute \( \langle \partial V(j, w'), w - w' \rangle \) in all three cases:

**Case** \( j = y \):

\[
\langle \partial V(y, w'), w - w' \rangle = -\phi'(1 - \langle c_y, w' \rangle) \langle c_y, w - w' \rangle = -\phi'(1 - \langle c_y, w' \rangle) \left( \frac{c_y - c_{y'}}{2}, w \right).
\]

37
Case $j = y'$:

$$\langle \partial V(y', w'), w - w' \rangle = -\phi'(1 - \langle c_{y'}, w' \rangle) \langle c_{y'}, w - w' \rangle = -\phi'(1 - \langle c_{y'}, w' \rangle) \left( \frac{c_{y'} - c_y}{2}, w \right).$$

Case $j \neq y, y'$:

$$\langle \partial V(j, w'), w - w' \rangle = -\phi'(1 - \langle c_j, w' \rangle) \langle c_j, w - w' \rangle = 0.$$

Let $\lambda_y(w') = \frac{1}{2} \phi'(1 - \langle c_y, w' \rangle)$, and $\lambda_{y'}(w') = \frac{1}{2} \phi'(1 - \langle c_{y'}, w' \rangle)$ hence:

$$\langle \partial \Psi_V(w', p), w - w' \rangle = (-p_y \lambda_y(w') + p_{y'} \lambda_{y'}(w')) \langle c_y - c_{y'}, w \rangle.$$

For (DC Margin), the proof is similar noting that:

$$\Psi_V(w, p) = \sum_{y \in \mathcal{Y}} p_y \sum_{j \neq y} \phi \left( \frac{1}{q - 1} + \langle c_j, w \rangle \right) = \sum_{y \in \mathcal{Y}} (1 - p_y) \phi \left( \frac{1}{q - 1} + \langle c_y, w \rangle \right).$$

For (AC Margin), the proof considers also three cases:

$$\partial \Psi_V(w, p) = \sum_{j \neq j} p_j \partial V(j, w), \ V(j, w) = \sum_{k \neq j} \phi \left( \frac{q}{q - 1} + \langle c_k - c_j, w \rangle \right).$$

Let $\phi'_{k,j}(w) = \phi' \left( \frac{q}{q - 1} + \langle c_k - c_j, w \rangle \right)$, note that $\phi'_{y',y}(w') = \phi'_{y,y'}(w') = \phi' \left( \frac{q}{q - 1} \right)$.

Case $j = y$:

$$\langle \partial V(y, w'), w - w' \rangle = \sum_{k \neq y} \phi'_{k,y}(w') \langle c_k - c_y, w - w' \rangle$$

$$= \sum_{k \neq y, y'} \phi'_{k,y}(w') \langle c_k - c_y, w - w' \rangle + \phi'_{y',y}(w') \langle c_{y'} - c_y, w - w' \rangle$$

$$= \sum_{k \neq y, y'} \phi'_{k,y}(w') \langle -c_y, w - w' \rangle + \phi' \left( \frac{q}{q - 1} \right) \langle c_{y'} - c_y, w \rangle$$

$$= \langle c_{y'} - c_y, w \rangle \left( \phi' \left( \frac{q}{q - 1} \right) + \frac{1}{2} \sum_{k \neq y, y'} \phi'_{k,y}(w') \right).$$

For the Case $j = y'$, a similar proof shows that:

$$\langle \partial V(y', w'), w - w' \rangle = -\langle c_{y'} - c_y, w \rangle \left( \phi' \left( \frac{q}{q - 1} \right) + \frac{1}{2} \sum_{k \neq y, y'} \phi'_{k,y}(w') \right).$$

38
Finally for $j \neq y, y'$:

$$
\langle \partial V(j, w'), w - w' \rangle = \sum_{k \neq j} \phi'_{k,j}(w') \langle c_k - c_j, w - w' \rangle
$$

$$
= \phi'_{y,j}(w') \langle c_y, w - w' \rangle + \phi'_{y',j}(w') \langle c_{y'}, w - w' \rangle
$$

$$
= \langle c_y - c_{y'}, w \rangle \frac{1}{2} \left( \phi'_{y,j}(w') - \phi'_{y',j}(w') \right).
$$

$\phi$ differentiable implies that $\lambda_y(w') = \lambda_{y'}(w')$, $\phi'_{y,j}(w') = \phi'_{y',j}(w')$, and $\langle \partial V(j, w'), w - w' \rangle = 0$. Hence:

$$
\langle \partial \Psi_V(w', p), w - w' \rangle = \langle c_y - c_{y'}, w \rangle \left( -\lambda_y(w')p_y + \lambda_{y'}(w')p_{y'} + \sum_{j \neq y, y'} \lambda_j(w')p_j \right).
$$

□
Chapter 4

Multiclass Classification with Simplex Support Vector Machines

In this chapter we study the consistency and the relaxation error of simplex Support Vector Machine (SVM). We also present the dual formulation of this problem and a stochastic gradient algorithm for linear simplex SVM.

4.1 Introduction

In Chapter 3 we studied the consistency of simplex multiclass loss functions under the three margin definitions (HS, DC, and AC). Interestingly we found that the simplex hinge loss, used in the Support Vector Machines (SVM), is classification calibrated for the DC margin.

In this chapter we turn our attention to DC-SVM, analyze its relaxation error, and derive practical algorithms for classification using that loss.

We explain in the following what we mean by the relaxation error analysis. As we discussed in Chapter 2, classification calibration or Fisher consistency ensures that the minimizer of the expected loss $\mathbb{E}_V(f)$, $f_\rho$ is consistent with the Bayes rule $b_\rho$, but we also highlighted the need of comparison inequalities that relate the excess expected $V$-loss of an arbitrary vector valued function $f$,

$$\Delta \mathcal{E}_V(f) = \mathcal{E}_V(f) - \mathcal{E}_V(f_\rho),$$

to the excess misclassification risk,

$$\Delta R(f) = R(D(f)) - R(D(f_\rho)).$$

Classification calibration allows us to replace $D(f_\rho)$ by $b_\rho$ and to write $\Delta R(f) = R(D(f)) - R(b_\rho)$. The relaxation error analysis consists in deriving explicit comparison inequalities of the form:

$$R(D(f)) - R(b_\rho) \leq \psi (\mathcal{E}_V(f) - \mathcal{E}_V(f_\rho)), \quad (4.1)$$

40
where $\psi$ is an increasing scalar function. Hence, we can assess the asymptotic convergence of a sequence of functions $f_N$, and obtain what is called Bayes consistency, i.e. if we have a sequence of function $f_N$, such that $\mathcal{E}_V(f_N) \to \mathcal{E}_V(f_\rho)$, then by virtue of equation (4.1) we have also $R(D(f_N)) \to R(b_\rho)$.

Until now our analysis was restricted to the expected risk that is indeed a theoretical object, since in practice we don’t have access to the distribution of the data $\rho$, and we have access to a training set $S = \{(x_i, y_i), x_i \in \mathcal{X}, y_i \in \mathcal{Y}, i = 1 \ldots N\}$, consisting of labeled data-points. While minimizing our expected risk, we were seeking for the minimum over the set of all measurable functions, which is not indeed possible in practice, hence we need to restrict our search over a set of functions $\mathcal{H}$ called the Hypothesis Class. Given a simplex loss function $V$, a training set $S$, and a hypothesis class $\mathcal{H}$ of vector valued functions, effective multiclass learning algorithms are obtained by doing the following Empirical Risk Minimization (ERM):

$$\min_{f \in \mathcal{H}} \hat{\mathcal{E}}_V(f) := \min_{f \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^{N} V(y_i, f(x_i)).$$

(4.2)

The minimization of the empirical risk (4.2), produces a function $\hat{f}_N$, and using tools from empirical processes theory we can assess the convergence of $\mathcal{E}_V(\hat{f}_N)$ to $\mathcal{E}_V(f_\rho)$. Naturally this convergence depends on two type of errors, the first is a statistical error and depends on the size of the data set $N$, the second is an estimation error, and depends on the complexity of the Hypothesis Class $\mathcal{H}$. Hence we can obtain bounds on $\Delta \mathcal{E}_V(\hat{f}_N)$, called generalization bounds, and by virtue of comparison inequalities of type (4.1) we obtain bounds on $\Delta R(\hat{f}_N)$.

The choice of a hypothesis class is crucial in order to ensure tractability and well-posedness of the empirical problem. In this chapter we consider regularization in Vector Valued Reproducing Kernel Hilbert Space (vv-RKHS) [MP04]. Let $\mathcal{H}_k$ be a vv-RKHS, the hypothesis class corresponds to a ball in the vv-RKHS

$$\mathcal{H} = \{f \in \mathcal{H}_k, \|f\|_{\mathcal{H}_k}^2 \leq R\},$$

where $\|f\|_{\mathcal{H}_k}^2$ is the norm in the vv-RKHS. An equivalent form of problem (4.2) in term of regularization is therefore:

$$\min_{f \in \mathcal{H}_k} \frac{1}{N} \sum_{i=1}^{N} V(y_i, f(x_i)) + \lambda \|f\|_{\mathcal{H}_k}^2,$$

where $\lambda$ is the regularization parameter ($\lambda > 0$).

This Chapter is organized as follows: in Section 4.2 we review the previous work on multiclass SVM. In Section 4.3 we study the consistency of DC-SVM and derive a comparison inequality relating the excess misclassification risk to the expected $V$-loss. In section 4.4 we review Vector Valued RKHS, and present generalization bounds for learning with DC-SVM in the framework of vv-RKHS, in Section 4.4.2. Finally we present the dual formulation of DC-
SVM in Section 5.4, as well as a Stochastic Gradient Descent (SGD) algorithm for solving linear DC-SVM.

4.2 Multiclass SVM: Previous Work

In [Liu07, VWR11], authors developed consistent multiclass SVM (M-SVM). Consider $q$ classes, the classification rule is found by applying a suitable prediction/decoding map to a function

$$f : \mathcal{X} \rightarrow \mathbb{R}^q$$

where $f$ is found considering the minimization of a loss function $V : \mathcal{Y} \times \mathbb{R}^q \rightarrow \mathbb{R}^+$, where

$$V(y, f(x)) = \sum_{j \neq y} \left\{ \frac{1}{q - 1} + f_j(x) \right\}_+.$$

In this setup, fisher consistency is obtained if the considered functions satisfy the so called sum to zero constraint [Liu07]

$$\sum_{y \in \mathcal{Y}} f^y(x) = 0, \text{ for all } x \in \mathcal{X}. \quad (4.3)$$

The latter requirement is problematic since it makes the computations in the corresponding algorithms more involved and is in fact often ignored, so that practical algorithms often come with no consistency guarantees. The main difference in our encoding is that the sum to zero constraint is automatically satisfied by virtue of properties of the simplex coding, for $f : \mathcal{X} \rightarrow \mathbb{R}^{q-1}$:

$$\sum_{y \in \mathcal{Y}} c_y = 0 \implies \sum_{y \in \mathcal{Y}} \langle c_y, f(x) \rangle = 0, \text{ for all } x \in \mathcal{X}.$$

Finally generalization bounds for M-SVM, under the sum to zero constraint were studied in [CX06]. In the following sections we study the consistency of DC-SVM and derive generalization bounds for a specific hypothesis class namely vv-RKHS.

4.3 DC-SVM Calibration and Comparison Inequality

Recall that the DC-SVM loss function is given by:

$$V(y, w) = \sum_{j \neq y} \left\{ \frac{1}{q - 1} + \langle c_j, w \rangle \right\}_+, \ y \in \mathcal{Y}. \quad (4.4)$$
We know from Chapter 3 that this loss is classification calibrated. Let $p \in \Delta _q^y$, and $w^*$ be the minimizer of the inner risk $\Psi_V(w,p)$, in Theorem 2 we show that:

$$w^* = c_y,$$  \hfill (4.5)

which is another confirmation of the classification calibration of this loss. Hence for $x \in \mathcal{X}$, we can define (up to $\rho_x$-null sets):

$$f_p(x) = c_{b_p(x)}, \quad b_p(x) = \arg \max_{j \in Y} \rho_j(x).$$  \hfill (4.6)

Observe that for a measurable function $f$ we have:

$$\mathcal{E}_V(f) - \mathcal{E}_V(f_p) = \int_{\mathcal{X}} (\Psi_V(f(x), \rho(x)) - \Psi_V(f_p(x), \rho(x))) d\rho_x(x)$$

$$R(D(f)) - R(b_p) = \int_{\mathcal{X}} (\rho_{b_p(x)} - \rho_{D(f(x))}) d\rho_x(x).$$  \hfill (4.7)

Hence to derive comparison inequalities as in (4.1), it is sufficient to derive bounds relating the excess inner risk, for $p \in \Delta _q^y$, and $w \in S_{y'}, y' \in \mathcal{Y}, y' \neq y$:

$$\Psi_V(w, p) - \Psi_V(w^*, p),$$ \hfill (4.8)

to the following probability discrepancy:

$$p_y - p_{y'},$$ \hfill (4.9)

and then integrate to get the final result. We call a comparison inequality of type (4.1), a comparison inequality with constant $K$, and exponent $\gamma$, if $\psi$ has the following form:

$$\psi(t) = Kt^\gamma, \ t > 0.$$  

Note that small constants and large exponent are desirable.

**Theorem 2 (Comparison Inequality).** Consider the DC-SVM loss function the following holds:

1. For $p \in \Delta _q^y$, $\Psi_V(., p)$ is minimized by $c_y$.

2. For $p \in \Delta _q^y$, $w \in S_{y'}$, $y' \in \mathcal{Y}, y' \neq y$ we have the following comparison inequality,

$$p_y - p_{y'} \leq (q - 1)(\Psi_V(w, p) - \Psi_V(c_y, p)).$$ \hfill (4.10)

3. For any measurable function $f$ we have:

$$R(D(f)) - R(D(f_p)) \leq (q - 1)(\mathcal{E}(f) - \mathcal{E}(f_p)),$$ \hfill (4.11)
where $f_p$ is given in equation (4.6).

Proof of Theorem 2. The proof of Theorem 2 is given in Section 4.7.

Theorem 2 shows that DC-SVM is indeed consistent and that we have an explicit comparison inequality, with constant $q - 1$, and exponent 1.

In order to have tractable algorithms, we need to consider a training set $S$ and a hypothesis class $\mathcal{H}$ of vector valued functions, and consider the following empirical Risk minimization:

$$\min_{f \in \mathcal{H}} \hat{\mathcal{E}}_V(f) := \min_{f \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^{N} V(y_i, f(x_i)).$$

(4.12)

In the next Section we study a particular hypothesis class, the so called vector valued RKHS (vv-RKHS).

4.4 Regularization in Vector Valued RKHS

4.4.1 Vector Valued RKHS (vv-RKHS)

As discussed in the introduction, the framework of vector valued Reproducing Kernel Hilbert Spaces (RKHSs) provides a natural choice for hypotheses spaces and regularizers in the multiclass setting. The definition of RKHS for vector valued functions parallels the one in the scalar case [Aro50], with the main difference that the reproducing kernel is now matrix valued – see [MP05, CDVT06] and references therein.

Let $\mathcal{X}$ be a set, a matrix valued reproducing kernel is a symmetric function $\Gamma : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{D \times D}$, such that for any $x, x' \in \mathcal{X}$, $\Gamma(x, x')$ is a positive semi-definite matrix, and $\sum_{i,j=1}^{N} (a_i, \Gamma(x_i, x_j) a_j) \geq 0$, for all $x_1, \ldots, x_N \in \mathcal{X}$ and $a_1, \ldots, a_N \in \mathbb{R}^D$.

Definition 7 (vv-RKHS, [MP05]). A vector valued RKHS is a Hilbert space $(\mathcal{H}, \langle \cdot , \cdot \rangle_{\mathcal{H}_\Gamma})$ of functions $f : \mathcal{X} \to \mathbb{R}^D$, such that for every $a \in \mathbb{R}^D$, and $x \in \mathcal{X}$, $\Gamma(x, \cdot) a$ belongs to $\mathcal{H}_\Gamma$, and moreover $\Gamma$ has the reproducing property

$$\langle f, \Gamma(x, \cdot) a \rangle_{\mathcal{H}_\Gamma} = \langle f(x), a \rangle,$$

where $\langle \cdot , \cdot \rangle$ is the dot product in $\mathbb{R}^D$.

Then, the choice of the kernel can be interpreted as inducing a representation for the functions of interest. Note that for $D = 1$ we recover the classic theory of scalar valued RKHS. Note that for our case $D = q - 1$, that is the dimension of the simplex coding. In the following we restrict our attention to kernels of the form

$$\Gamma(x, x') = k(x, x') B, \ B \in \mathbb{R}^{D \times D}, \ B = I_D,$$

(4.13)
where $k: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a scalar valued reproducing kernel. In the following we will discuss in particular the case where the kernel is induced by a finite dimensional feature map,

$$k(x, x) = \langle \Phi(x), \Phi(x') \rangle,$$

where $\Phi: \mathcal{X} \to \mathbb{R}^p$, (4.14)

and $\langle \cdot, \cdot \rangle$ is the inner product in $\mathbb{R}^p$. In this case we can write each function in $\mathcal{H}$ as

$$f(x) = W^T \Phi(x),$$

where $W \in \mathbb{R}^{p \times D}$.

An important property of vv-RKHS is the representer theorem [MP05]. In particular if we consider the regularized empirical risk minimization,

$$\min_{f \in \mathcal{H}_r} \frac{1}{N} \sum_{i=1}^N V(y_i, f(x_i)) + \lambda \|f\|^2_{\mathcal{H}_r},$$

where $V$ is a convex loss in the second argument, $\lambda > 0$ is a regularization parameter, and $\|f\|^2_{\mathcal{H}_r} = \langle f, f \rangle_{\mathcal{H}_r}$, is the norm in the vv-RKHS, then the solution of this minimization problem has the following form:

$$f^*_N(x) = \sum_{i=1}^N \Gamma(x, x_i)a_i,$$

(4.15)

for some $a_i \in \mathbb{R}^D, i = 1 \ldots N$, and $\|f^*_N\|^2_{\mathcal{H}_r} = \sum_{i=1}^N \langle a_i, \Gamma(x_i, x_i)a_j \rangle$.

### 4.4.2 Generalization Bounds

We turn now to the regularized ERM for DC-SVM, and study its relaxation error analysis by mean of generalization bounds on the excess misclassification risk. Recall that the empirical risk is:

$$\hat{\mathcal{E}}_V(f) = \frac{1}{N} \sum_{i=1}^N V(y_i, f(x_i)).$$

Regularized ERM corresponds to the following problem:

$$\min_{f \in \mathcal{H}_r} \hat{\mathcal{E}}_V(f) + \lambda \|f\|^2_{\mathcal{H}_r},$$

(4.16)

where $V$ here corresponds to DC-SVM loss. Let $f^*_N$, be a minimizer of the regularized empirical empirical risk $\hat{\mathcal{E}}_V(f) + \lambda \|f\|^2_{\mathcal{H}_r}$ in $\mathcal{H}_r$. Consider also the regularized expected risk minimization:

$$\min_{f \in \mathcal{H}_r} \mathcal{E}_V(f) + \lambda \|f\|^2_{\mathcal{H}_r},$$

(4.17)

and let $f^\lambda$ be a minimizer of $\mathcal{E}_V(f) + \lambda \|f\|^2_{\mathcal{H}_r}$ in $\mathcal{H}_r$. 45
Theorem 3 (Generalization Bound). Let $\mathcal{H}_\Gamma$ be a vv-RKHS induced by the kernel $\Gamma$, and let $\Gamma_{\text{max}} = \sup_{x \in \mathcal{X}} \|\Gamma(x, x)\|$. Assume also that the kernel is universal such that:

$$\mathcal{E}(f_\rho) = \inf_{f \in \mathcal{H}_\Gamma} \mathcal{E}(f).$$

Let $\hat{f}_N^{\lambda}$ be a minimizer of the regularized empirical empirical risk $\hat{\mathcal{E}}_V(f) + \lambda \|f\|_{\mathcal{H}_\Gamma}^2$ in $\mathcal{H}_\Gamma$, where $V$ is the DC-SVM loss, $\lambda$ is the regularization parameter $0 < \lambda < 1$. Assume there exist $c > 0$, $\beta \in [0, 1]$, such that the approximation error:

$$\mathcal{A}(\lambda) = \inf_{f \in \mathcal{H}_\Gamma} (\mathcal{E}_V(f) + \lambda \|f\|_{\mathcal{H}_\Gamma}^2 - \mathcal{E}_V(f_\rho)) \leq c\lambda^\beta,$$

then $\hat{f}_N^{\lambda}$ satisfies:

$$R(D(\hat{f}_N^{\lambda})) - R(D(f_\rho)) \leq 4(q)(q-1)\sqrt{\frac{\Gamma_{\text{max}}}{N\lambda}} + 2 \left(\frac{(q-1)^2}{\lambda} \sqrt{\frac{\Gamma_{\text{max}}}{\lambda}} + (q-1)\right) \sqrt{\frac{2\ln(2/\delta)}{N}} + c(q-1)\lambda^\beta,$$

with probability $1 - \delta$, $\forall \delta > 0$.

Proof of Theorem 3. The proof of Theorem 3 is given in Section 4.7. \qed

Theorem 3 shows that the excess misclassification error is indeed decreasing in the sample size number $N$.

4.5 Algorithms

In this section, we derive the dual formulation for DC-SVM and an online algorithm for solving linear DC-SVM in the primal. As well as an algorithm for computing the simplex code and show its correctness.

4.5.1 Dual Formulation

In this section we restrict our attention to the vv-RKHS induced by separable kernels of the form (4.13), and we note $\mathcal{H}_\Gamma$ simply by $\mathcal{H}$. By virtue of the representer theorem the minimizer of (4.16) satisfies:

$$\hat{f}_N^{\lambda}(x) = \sum_{i=1}^{N} K(x, x_i)a_i, \ a_i \in \mathbb{R}^{q-1}, \ i = 1 \ldots N,$$

and $\|\hat{f}_N^{\lambda}\|_{\mathcal{H}}^2 = \sum_{i=1}^{N} K(x_i, x_j)\langle a_i, a_j \rangle$.

Following the notation for binary SVM, we penalize the loss instead of regularizing the norm.
for convenience, and study the dual formulation of the following problem:

$$\min_{f \in \mathcal{H}} \left\{ C_0 \sum_{i=1}^{N} \sum_{y \neq y_i} \left| \langle c_y, f(x_i) \rangle \right| + \frac{1}{q-1} + \frac{1}{2} \| f \|^2_{\mathcal{H}} \right\}$$  \hspace{1cm} (4.19)$$

where \( C_0 = \frac{1}{2N}. \)

**Proposition 1** (Dual Formulation). For the DC-SVM the coefficients in the representer theorem are given by

$$a_i = - \sum_{y \neq y_i} \alpha_i^y c_y, \quad i = 1, \ldots, N,$$

where \( \alpha_i = (\alpha_i^y)_{y \in \mathcal{Y}} \in \mathbb{R}^q, i = 1, \ldots, N, \) solve the following quadratic programming (QP) problem

$$\max_{a_1, \ldots, a_n \in \mathbb{R}^q} \left\{ \frac{1}{2} \sum_{y \neq y', i, j} \alpha_i^y K_{ij} G_{yy'} \alpha_j^{y'} + \frac{1}{q-1} \sum_{i=1}^{N} \sum_{y \in \mathcal{Y}} \alpha_i^y \right\}$$ \hspace{1cm} (4.20)$$

subject to \( 0 \leq \alpha_i^y \leq C_0(1 - \delta_{y,y_i}), \quad \forall i = 1, \ldots, N, y \in \mathcal{Y}. \)

where \( G_{yy'} = \langle c_y, c_{y'} \rangle \forall y, y' \in \mathcal{Y} \) and \( C_0 = \frac{1}{2N}, \) \( \alpha_i = (\alpha_i^y)_{y \in \mathcal{Y}} \in \mathbb{R}^q, \) for \( i = 1, \ldots, N \) and \( \delta_{i,j} = 1 \) if \( i = j, \) and 0 elsewhere.

**Proof of Proposition 1.** The proof of Proposition 1 is given in Section 4.7. \( \square \)

### 4.5.2 Primal Solution with Stochastic Gradient

The regularized estimators induced by the simplex loss functions can be computed by mean of online/incremental first order (sub) gradient methods. Indeed, when considering finite dimensional feature maps \( \Phi : \mathcal{X} \to \mathbb{R}^p, \) these strategies offer computationally feasible solutions to train estimators for large datasets where neither a \( p \) by \( p \) or an \( N \) by \( N \) matrix fit in memory. In that case the linear DC-SVM primal problem has the following form

$$\min_{W \in \mathbb{R}^{p \times (q-1)}} \frac{1}{N} \sum_{i=1}^{N} V(y_i, W^T \Phi(x_i)) + \lambda\|W\|^2_F,$$

where \( \| \cdot \|_F \) corresponds to the Frobenius norm. Following [SSSS07] we can alternate a step of stochastic descent on a data point:

$$W_{\text{tmp}} = (1 - \eta_i \lambda)W_i - \eta_i \partial(V(y_i, W_i^T \Phi(x_i)))$$

and a projection on the Frobenius ball

$$W_i = \min \left( 1, \frac{1}{\sqrt{\lambda \|W_{\text{tmp}}\|_F}} \right) W_{\text{tmp}}.$$

47
For DC-SVM loss we have

$$\partial(V(y_i, W_i \Phi(x_i))) = \Phi(x_i) \left( \sum_{k \in I_i} c_k \right)^\top,$$

where $I_i = \{y \neq y_i | \langle c_y, W^\top \Phi(x_i) \rangle > -\frac{1}{q-1} \}$.

**Algorithm 1** Multiclass Pegasos

**INPUT:** $S, \lambda, L$

**INITIALIZE:** $W_0 = 0$

FOR $i = 1 \ldots L$

$\eta_i = \frac{1}{\lambda_i}$

$W_{\text{tmp}} = (1 - \eta_i \lambda)W_i - \eta_i \partial(V(y_i, W_i^\top \Phi(x_i)))$

$W_i = \min(1, \frac{1}{\sqrt{\lambda_i W_{\text{tmp}}}})W_{\text{tmp}}$

**OUTPUT:** $W_L$

### 4.5.3 Computing the Simplex Coding

In this section we give a recursive algorithm that produces the simplex coding vectors in $\mathbb{R}^{q-1}$.

**Algorithm 2** Simplex Code

**SET:** $C[2] = [1 - 1]$,

FOR $i = 2, \ldots, q - 1$

$u = (-\frac{1}{i} \ldots - \frac{1}{i})$ (column vector in $\mathbb{R}^i$)

$v = (0, \ldots, 0)$ (column vector in $\mathbb{R}^{i-1}$)

$C[i + 1] = \begin{pmatrix} 1 & u^T \\ v & C[i] \times \sqrt{1 - \frac{1}{i^2}} \end{pmatrix}$

ENDFOR

**OUTPUT:** $C[q]$

**Lemma 3.** The $q$ columns of $C[q]$ are a set of $q - 1$ dimensional vectors satisfying the properties of the simplex coding.

**Proof.** The statement is proved by induction. The base case is trivially true. Let $c_1 \ldots c_i$ be the columns of $C[i]$ and $b_1 \ldots b_{i+1}$ be the columns of $C[i+1]$. By construction $b_1 = (1, 0, \ldots, 0)$ and $b_m = (-\frac{1}{i}, \sqrt{1 - \frac{1}{i^2}} c_m)$ for all $m = 2, \ldots, i + 1$.

Assume $C[i]$ to satisfy definition 1, that is $\|c_m\|^2 = 1$, for $1 \leq m \leq i$ and $\langle c_m, c_n \rangle = -\frac{1}{i-1} \forall m \neq n$. Indeed, a direct calculation shows that $\|b_1\| = 1$ and $\langle b_1, b_m \rangle = -\frac{1}{i}, \forall m \neq 1$. 

48
Moreover, for $m \neq n$ such that $2 \leq m, n \leq i + 1$ we have:

\[
\|b_m\|^2 = \frac{1}{i^2} + \left(1 - \frac{1}{i^2}\right) \|c_m\|^2 = 1,
\]

and

\[
\langle b_m, b_n \rangle = \frac{1}{i^2} + \left(1 - \frac{1}{i^2}\right) \langle c_m, c_n \rangle = \frac{1}{i^2} - \left(1 - \frac{1}{i^2}\right) \frac{1}{i - 1} = \frac{1}{i^2} - \frac{i + 1}{i} = -\frac{1}{i}.
\]

\[\square\]

### 4.6 Conclusion

In this Chapter we studied the consistency and the relaxation error analysis for DC-SVM, and derived the dual form as well as an incremental algorithm for solving the primal linear formulation. The following Chapter answer the following question:

*What about classification with Simplex Least Squares? Is it classification calibrated? How does it compare to DC-SVM?*

### 4.7 Proofs of Theorems 2 and 3 and Proposition 1

**Proof of Theorem 2.** 1) For $p \in \Delta_{q}^y$, in order to prove that $c_y$, achieves the minimum of $\Psi_V(w, p)$. Since $\Psi_V(., p)$ is convex by the convexity of $V$, it is sufficient to show that

\[0 \in \partial \Psi_V(c_y, p),\]

where $\partial \Psi_V(c_y, p)$, is the subgradient of $\Psi_V(., p)$ at $c_y$. Note that:

\[
\Psi_V(w, p) = \sum_{j \in \mathcal{Y}} p_j \sum_{k \neq j} \left| \frac{1}{q - 1} + \langle c_k, w \rangle \right|_+ = \sum_{j \in \mathcal{Y}} (1 - p_j) \left| \frac{1}{q - 1} + \langle c_j, w \rangle \right|_+.
\]

It follows that:

\[
\partial \Psi_V(w, p) = \sum_{j \in \mathcal{Y}} (1 - p_j) \partial \left| \frac{1}{q - 1} + \langle c_j, w \rangle \right|_+. \tag{4.21}
\]
Note that for $j \in \mathcal{Y}$, we have:
\[\partial \left| \frac{1}{q-1} + \langle c_j, w \rangle \right|_+ = \begin{cases} 
    c_j & \text{if } \langle c_j, w \rangle > -\frac{1}{q-1} \\
    \lambda_j c_j, 0 < \lambda_j < 1 & \text{if } \langle w, c_j \rangle = -\frac{1}{q-1} \\
    0 & \text{if } \langle w, c_j \rangle < -\frac{1}{q-1}.
\end{cases}\]

Hence the subgradient evaluated at $w = c_y$ is:
\[\partial \left| \frac{1}{q-1} + \langle c_j, w \rangle \right|_+ = \begin{cases} 
    \lambda_j c_j, 0 < \lambda_j < 1 & \text{if } j \neq y \\
    c_y & \text{if } j = y
\end{cases}\]

Hence the subgradient of $\Psi_V(., p)$ at $c_y$ has the following form:
\[\partial \Psi_V(c_y, p) = \left\{ v \in \mathbb{R}^{q-1} \mid v = (1 - p_y)c_y + \sum_{j \in \mathcal{Y}, j \neq y} (1 - p_j)\lambda_j c_j, 0 < \lambda_j < 1, j \in \mathcal{Y}, j \neq y \right\}.\] (4.22)

Recall that $p \in \Delta_y^q$, and therefore:
\[p_y > p_j, \forall j \in \mathcal{Y}/\{y\}.\] (4.23)

We are now ready to show that $0 \in \partial \Psi_V(c_y, p)$. Consider
\[\lambda_j^* = \frac{1 - p_y}{1 - p_j}, \forall j \in \mathcal{Y}/\{y\},\]
by equation (4.23), we have $0 < \lambda_j^* < 1, \forall j \in \mathcal{Y}/\{y\}$. Setting $\lambda_j = \lambda_j^*$ in (4.22) we have finally:
\[v^* = (1 - p_y)c_y + \sum_{j \neq y} (1 - p_j)\lambda_j^* c_j = (1 - p_y)c_y + (1 - p_y)\sum_{j \neq y} c_j\]
\[= (1 - p_y)c_y + (1 - p_y)(-c_y)\]
\[= 0,
\]
where we used the property of the simplex coding that $\sum_{j \in \mathcal{Y}} c_j = 0$. Hence $0 \in \partial \Psi_V(c_y, p)$, and $c_y$ achieves the minimum of $\Psi_V(., p)$ for $p \in \Delta_y^q$.

2) Let $p \in \Delta_y^q$, and let $w \in S_y'$, $y' \in \mathcal{Y}, y' \neq y$. Let $I = \{ i \in \mathcal{Y} \mid \langle c_i, w \rangle \geq -\frac{1}{q-1} \}$. $y'$ belongs to the set $I$, as the maximum of $\langle c_j, w \rangle$ is achieved for $j = y'$, and the equation $\sum_{j \in \mathcal{Y}} \langle c_j, w \rangle = 0$ is always satisfied so that $\langle c_{y'}, w \rangle$ is necessarily positive. Note that for any
$w \in \mathbb{R}^{q-1}$, we have:
\[
\Psi_V(c_y, p) = (1 - p_y) \frac{q}{q - 1} = (1 - p_y) \sum_{j \in \mathcal{Y}} \left( \frac{1}{q - 1} + \langle c_j, w \rangle \right)
\]
\[
\Psi_V(w, p) - \Psi_V(c_y, p) = \sum_{j \in \mathcal{Y}} (1 - p_j) \left| \langle c_j, w \rangle + \frac{1}{q - 1} \right| - (1 - p_y) \sum_{j \in \mathcal{Y}} \left( \frac{1}{q - 1} + \langle c_j, w \rangle \right)
\]
\[
= \sum_{i \in I} (p_y - p_i) \left( \langle c_i, w \rangle + \frac{1}{q - 1} \right) - (1 - p_y) \sum_{i \not\in I} \left( \langle c_i, w \rangle + \frac{1}{q - 1} \right)
\]
\[
= \frac{1}{q - 1} (p_y - p_{y'}) + \left( \frac{p_y - p_{y'}}{q - 1} \langle c_{y'}, w \rangle \right) + \sum_{i \in I, i \not= y'} (p_y - p_i) \left( \langle c_i, w \rangle + \frac{1}{q - 1} \right) \geq 0, \text{ by definition of } y \text{ and } w
\]
\[
- (1 - p_y) \sum_{i \in I} \left( \langle c_i, w \rangle + \frac{1}{q - 1} \right)
\]
\[
\geq \frac{1}{q - 1} (p_y - p_{y'}).
\]

The last inequality is due to the positivity of the other terms.

3) This follows simply from combining (4.10), and (4.7). Let
\[
\mathcal{X}_f = \{ x \in \mathcal{X}, D(f(x)) \neq b_\rho(x) \},
\]
\[
R(D(f)) - R(D(f_\rho)) = \int_{\mathcal{X}} (\rho_{b_\rho(x)}(x) - \rho_{D(f(x))}(x)) \, d\rho_{\mathcal{X}}(x)
\]
\[
= \int_{\mathcal{X}_f} (\rho_{b_\rho(x)}(x) - \rho_{D(f(x))}(x)) \, d\rho_{\mathcal{X}}(x)
\]
\[
\leq (q - 1) \int_{\mathcal{X}_f} (\Psi_V(f(x), \rho(x)) - \Psi_V(f_\rho(x), \rho(x))) \, d\rho_{\mathcal{X}}(x)
\]
\[
\leq (q - 1) \int_{\mathcal{X}} (\Psi_V(f(x), \rho(x)) - \Psi_V(f_\rho(x), \rho(x))) \, d\rho_{\mathcal{X}}(x),
\]
the last inequality follows from the definition of $f_\rho$, as $\Psi_V(f(x), \rho(x)) \geq \Psi_V(f_\rho(x), \rho(x))$. □

Proof of Theorem 3. We decompose the excess error $\mathcal{E}_V(f_N) - \mathcal{E}_V(f_\rho)$, following standard
techniques as:

\[ \mathcal{E}_V(f_N^\lambda) - \mathcal{E}_V(f_\rho) = \left( \mathcal{E}_V(f_N^\lambda) - \hat{\mathcal{E}}_V(f_N^\lambda) \right) + \left( \left( \hat{\mathcal{E}}_V(f_N^\lambda) + \lambda \| f_N^\lambda \|^2_{\mathcal{H}_r} \right) - \left( \hat{\mathcal{E}}_V(f^\lambda) + \lambda \| f^\lambda \|^2_{\mathcal{H}_r} \right) \right) \\
+ \left( \hat{\mathcal{E}}_V(f^\lambda) - \mathcal{E}_V(f^\lambda) \right) + \left( \mathcal{E}_V(f^\lambda) - \mathcal{E}_V(f_\rho) + \lambda \| f^\lambda \|^2_{\mathcal{H}_r} \right) - \lambda \| f^\lambda \|^2_{\mathcal{H}_r}. \]

(4.28)

By the definition of \( \hat{f}_N^\lambda \) we have:

\[ \left( \hat{\mathcal{E}}_V(f_N^\lambda) + \lambda \| f_N^\lambda \|^2_{\mathcal{H}_r} \right) - \left( \hat{\mathcal{E}}_V(f^\lambda) + \lambda \| f^\lambda \|^2_{\mathcal{H}_r} \right) \leq 0. \]

It follows that:

\[ \mathcal{E}_V(f_N^\lambda) - \mathcal{E}_V(f_\rho) \leq \left( \mathcal{E}_V(f_N^\lambda) - \hat{\mathcal{E}}_V(f_N^\lambda) \right) + \left( \hat{\mathcal{E}}_V(f^\lambda) - \mathcal{E}_V(f^\lambda) \right) \\
+ \left( \mathcal{E}_V(f^\lambda) - \mathcal{E}_V(f_\rho) + \lambda \| f^\lambda \|^2_{\mathcal{H}_r} \right) = \left( \mathcal{E}_V(f_N^\lambda) - \hat{\mathcal{E}}_V(f_N^\lambda) \right) + \left( \hat{\mathcal{E}}_V(f^\lambda) - \mathcal{E}_V(f^\lambda) \right) \\
+ \inf_{f \in \mathcal{H}_r} \left( \mathcal{E}_V(f) + \lambda \| f \|^2_{\mathcal{H}_r} - \mathcal{E}_V(f_\rho) \right) \leq 2 \sup_{f \in \mathcal{H}_r} \left( \hat{\mathcal{E}}_V(f) - \mathcal{E}_V(f) \right) + \inf_{f \in \mathcal{H}_r} \left( \mathcal{E}_V(f) + \lambda \| f \|^2_{\mathcal{H}_r} - \mathcal{E}_V(f_\rho) \right). \]

(4.29)

where

\[ \mathcal{H}_{\Gamma,R} = \{ f \in \mathcal{H}_r, \| f \|_{\mathcal{H}_r} \leq R \}, \]

and \( R \) is a bound on \( \| f^\lambda \|_{\mathcal{H}_r}, \) and \( \| \hat{f}_N^\lambda \|_{\mathcal{H}_r}. \) By the positivity of the loss and the definition of \( f^\lambda, \) we have:

\[ \lambda \| f_\rho \|^2_{\mathcal{H}_r} \leq \mathcal{E}_V(f^\lambda) + \lambda \| f^\lambda \|^2_{\mathcal{H}_r} \leq \mathcal{E}_V(0) + 0 = 1, \]

a similar argument holds for \( \hat{f}_N^\lambda, \) and we conclude that \( \| f^\lambda \|_{\mathcal{H}_r} \leq \sqrt{\frac{1}{\lambda}}, \) and \( \| \hat{f}_N^\lambda \|_{\mathcal{H}_r} \leq \sqrt{\frac{1}{\lambda}}. \) Hence we set:

\[ R = \sqrt{\frac{1}{\lambda}}. \]

(4.30)

Let \( \epsilon_i \in \{-1, 1\}, i = 1 \ldots N, \) be \( N \) rademacher random variables with probability i.e \( \mathbb{P}(\epsilon_i = 1) = \mathbb{P}(\epsilon_i = -1) = \frac{1}{2}. \)

Let \( \ell(z) = \left| \frac{1}{q-1} + z \right|_+. \) We start by bounding the expectation of

\[ \mathcal{R}(\mathcal{H}_{\Gamma,R}) = \sup_{f \in \mathcal{H}_{\Gamma,R}} \left| \hat{\mathcal{E}}_V(f) - \mathcal{E}_V(f) \right|, \]
by a standard symmetrization argument [LT91] we have the Rademacher average:

\[
\mathbb{E}(\mathcal{R}(\mathcal{H}_{r,R})) \leq 2\mathbb{E}\left(\sup_{f \in \mathcal{H}_{r,R}} \frac{1}{N} \left| \sum_{i=1}^{N} \epsilon_i \sum_{j \neq y_i} (\ell((c_j, f(x_i))) - \ell(0)) \right| \right)
\]

\[
\leq 2 \sum_{j=0}^{q-1} \mathbb{E}\left(\sup_{f \in \mathcal{H}_{r,R}} \frac{1}{N} \left| \sum_{i=1}^{N} \epsilon_i (\ell((c_j, f(x_i))) - \ell(0)) \right| \right). \quad (4.31)
\]

Let \( \psi(z) = \ell(z) - \ell(0) \), \( \psi \) is a contraction with Lipchitz constant 1, we appeal now to the principle of contraction [LT91] and conclude that:

\[
\mathbb{E}\left(\sup_{f \in \mathcal{H}_{r,R}} \frac{1}{N} \left| \sum_{i=1}^{N} \epsilon_i (\ell((c_j, f(x_i))) - \ell(0)) \right| \right) \leq 2\mathbb{E}\left(\sup_{f \in \mathcal{H}_{r,R}} \frac{1}{N} \left| \sum_{i=1}^{N} \epsilon_i (c_j, f(x_i)) \right| \right)
\]

\[
= 2\mathbb{E}\left(\sup_{f \in \mathcal{H}_{r,R}} \left\langle f, \frac{1}{N} \sum_{i=1}^{N} \Gamma(x_i, \cdot)\epsilon_i c_j \right\rangle \right)
\]

\[
\leq \frac{2}{N} \sup_{f \in \mathcal{H}_{r,R}} \|f\|_{\mathcal{H}_r} \mathbb{E}\left(\left\| \sum_{i=1}^{N} \Gamma(x_i, \cdot)\epsilon_i c_j \right\|_{\mathcal{H}_r}^2 \right)^{\frac{1}{2}} \quad (4.32)
\]

On the other hand since \( \mathbb{E}(\epsilon_i^2) = 1 \), and \( \mathbb{E}(\epsilon_i \epsilon_j) = 0 \), \( i \neq j \), we have:

\[
\mathbb{E}\left(\left\| \sum_{i=1}^{N} \Gamma(x_i, \cdot)\epsilon_i c_j \right\|_{\mathcal{H}_r}^2 \right) = \mathbb{E}\sum_{i,j} \epsilon_i \epsilon_j \langle c_i, \Gamma(x_i, x_j) c_j \rangle = \sum_i \langle c_i, \Gamma(x_i, x_i) c_i \rangle
\]

\[
\leq N \max_i \|\Gamma(x_i, x_i)\| \leq N \Gamma_{\max}. \quad (4.33)
\]

Putting together (4.30), (4.31), (4.32), and (10.36), we obtain finally:

\[
\mathbb{E}(\mathcal{R}(\mathcal{H}_r)) \leq 2q \sqrt{\frac{\Gamma_{\max}}{N \lambda}}.
\]

**Lemma 4 (Hoeffding Inequality).** For any fixed random variable \( \xi \) on \( \mathcal{Z} \) satisfying:

\[
|\xi(z) - \mathbb{E}(\xi)| \leq B,
\]

we have with probability \( 1 - \delta \):

\[
\left| \frac{1}{N} \sum_{i=1}^{N} \xi(z_i) - \mathbb{E}(\xi) \right| \leq B \sqrt{\frac{2 \ln(2/\delta)}{N}}, \quad \forall \delta > 0. \quad (4.34)
\]
For \( f \in \mathcal{H}_{\Gamma,R} \), we have:

\[
|V(y_i, f(x_i)) - \mathbb{E}(V(y, f(x)))| \leq B = (q - 1) \left( \sqrt{\Gamma_{\text{max}} R + \frac{1}{q - 1}} \right).
\]

Let \( \delta > 0 \). Using hoeffding inequality it is possible now to bound the \( \mathcal{R}(\mathcal{H}_{\Gamma,R}) \), with a probability \( 1 - \delta \) by:

\[
\mathcal{R}(\mathcal{H}_{\Gamma,R}) \leq \mathbb{E}(\mathcal{R}(\mathcal{H}_{\Gamma})) + B \sqrt{\frac{2 \ln(2/\delta)}{N}} \\
\leq 2q \sqrt{\frac{\Gamma_{\text{max}}}{N \lambda}} + \left( (q - 1) \sqrt{\frac{\Gamma_{\text{max}}}{\lambda}} + 1 \right) \sqrt{\frac{2 \ln(2/\delta)}{N}}.
\] (4.35)

Let

\[
\mathcal{A}(\lambda) = \inf_{f \in \mathcal{H}_{\Gamma,R}} \left( \mathcal{E}_V(f) + \lambda \|f\|_{\mathcal{H}_R}^2 - \mathcal{E}_V(f_0) \right).
\]

The following lemma is an extract of Lemma 5.15 in [SC08].

**Lemma 5 ([SC08]).** The approximation error is an increasing, concave and continuous function of \( \mathcal{A} : [0, \infty) \to [0, \infty) \). Moreover, \( \mathcal{A}(0) = 0 \), and if it exists \( h : [0, 1] \to [0, \infty) \) such that \( \lim_{\lambda \to 0^+} h(\lambda) = 0 \) and \( \mathcal{A}(\lambda) \leq \lambda h(\lambda) \) for all \( \lambda \in [0, 1] \), then \( \mathcal{A}(\lambda) = 0 \) for all \( \lambda \geq 0 \).

We can assume that there exists \( c > 0, \beta \in [0, 1] \) such that:

\[
\mathcal{A}(\lambda) \leq c \lambda^\beta. \quad (4.36)
\]

Putting together equations (4.29), (4.35), and (4.36), we have finally, with probability \( 1 - \delta \):

\[
\mathcal{E}_V(f_N^\lambda) - \mathcal{E}_V(f_0) \leq 4q \sqrt{\frac{\Gamma_{\text{max}}}{N \lambda}} + 2 \left( (q - 1) \sqrt{\frac{\Gamma_{\text{max}}}{\lambda}} + 1 \right) \sqrt{\frac{2 \ln(2/\delta)}{N}} + c \lambda^\beta. \quad (4.37)
\]

It follows from the comparison inequality given in (4.11), that the following holds:

\[
R(D(f_N^\lambda)) - R(D(f_0)) \leq 4q(q - 1) \sqrt{\frac{\Gamma_{\text{max}}}{N \lambda}} + 2 \left( (q - 1)^2 \sqrt{\frac{\Gamma_{\text{max}}}{\lambda}} + (q - 1) \right) \sqrt{\frac{2 \ln(2/\delta)}{N}} + c(q - 1) \lambda^\beta, \quad (4.38)
\]

with probability \( 1 - \delta \).

\( \Box \)

**Proof of Proposition 1.** Following the notation for binary SVM, we penalize the loss instead of regularizing the loss for convenience:

\[
\min_{f \in \mathcal{H}} \left\{ C_0 \sum_{i=1}^{N} \sum_{y \neq y_i} |c_y, f(x_i)| + \frac{1}{q - 1} + \frac{1}{2} \|f\|_{\mathcal{H}}^2 \right\}
\]
where \( C_0 = \frac{1}{2N\lambda} \). Using the representer theorem (4.15), and introducing the slack variables \( \xi_i = (\xi_i^y)_{y \in \mathcal{Y}} \in \mathbb{R}^q \), for \( i = 1, \ldots, N \), we can write the above problem as,

\[
\min_{a_1, \ldots, a_N, \xi_1, \ldots, \xi_N \in \mathbb{R}^q} \left\{ C_0 \sum_{i=1}^{N} \sum_{y \neq y_i} \xi_i^y + \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} K_{ij} \langle a_i, a_j \rangle \right\}
\]

\[
\frac{1}{q-1} + \sum_{j=1}^{N} K_{ij} \langle a_j, c_y \rangle \leq \xi_i^y, \quad \forall \ i = 1 \ldots N, \text{ and } y \neq y_i,
\]

\[
\xi_i^y \geq 0, \quad \forall \ i = 1 \ldots N, \text{ and } y \neq y_i.
\]

Let \( \Xi = \{ \xi_i = (\xi_i^y)_{y \in \mathcal{Y}} \in \mathbb{R}^q, \ i = 1, \ldots, N \} \), \( \alpha = \{ \alpha_i = (\alpha_i^y)_{y \in \mathcal{Y}} \in \mathbb{R}^q, \ i = 1, \ldots, N \} \), \( \nu = \{ \nu_i = (\nu_i^y)_{y \in \mathcal{Y}} \in \mathbb{R}^q, \ i = 1, \ldots, N \} \) and consider the Lagrangian corresponding to the above problem given by,

\[
\mathcal{L}(A, \Xi, \alpha, \nu) = C_0 \sum_{i=1}^{N} \sum_{y \neq y_i} \xi_i^y + \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} K_{ij} \langle a_i, a_j \rangle + \sum_{i=1}^{N} \sum_{y \neq y_i} \alpha_i^y \left( \frac{1}{q-1} + \sum_{j=1}^{N} K_{ij} \langle a_j, c_y \rangle - \xi_i^y \right)
\]

\[
\sum_{i=1}^{N} \sum_{y \neq y_i} \nu_i^y
\]

\[
= \frac{1}{q-1} \sum_{i=1}^{N} \sum_{y \neq y_i} \alpha_i^y + \sum_{i=1}^{N} \sum_{y \neq y_i} (C_0 - \alpha_i^y - \nu_i^y) \xi_i^y + \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} K_{ij} \langle a_i, a_j \rangle
\]

\[
+ \sum_{i=1}^{N} \sum_{j=1}^{N} K_{ij} \left( \sum_{y \neq y_i} \alpha_i^y c_y, a_j \right)
\]

Considering the first order condition of optimality, a direct computation gives:

\[
a_i = -\sum_{y \neq y_i} \alpha_i^y c_y, \quad \text{with} \quad 0 \leq \alpha_i^y \leq C_0, \quad \forall \ i = 1 \ldots N, \text{ and } y \neq y_i. \tag{4.39}
\]

Let \( \alpha_i \) a \( q \) dimensional vector, with \( \alpha_i^{y_i} = 0 \), and \( \alpha = (\alpha_1, \ldots, \alpha_n) \) a \( n \times q \) dimensional vector. In the following we let \( \alpha_i^{y_i} \) free in the objective and add a constraint such that \( \alpha_i^{y_i} = 0, \quad \forall i \).

Recalling the definition of the matrices \( G \) and \( K \), the dual problem corresponding to (4.39) is given by, the maximization of

\[
L_D(\alpha) = -\frac{1}{2} \sum_{i,j} K_{ij} \sum_{y \neq y_i, y' \neq y_j} \alpha_i^y G_{yy'} \alpha_j^{y'} + \frac{1}{q-1} \sum_{i=1}^{N} \sum_{y \neq y_i} \alpha_i^y
\]

\[
= -\frac{1}{2} \sum_{y, y', i, j} \alpha_i^y K_{ij} G_{yy'} \alpha_j^{y'} + \frac{1}{q-1} \sum_{i=1}^{N} \sum_{y} \alpha_i^y
\]

55
subject to the constraints (4.39). Let $H_{ijyy'} = K_{ij}G_{yy'}$, $H$ is a $(N \times q) \times (N \times q)$ matrix. So that we can write:

$$L_D(\alpha) = -\frac{1}{2} \alpha^\top H \alpha + \frac{1}{q-1} 1_{N \times q}^\top \alpha$$

(4.40)

$$\max_{\alpha} -\frac{1}{2} \alpha^\top H \alpha + \frac{1}{q-1} 1_{N \times q}^\top \alpha$$

$$0 \leq \alpha_i^y \leq C_0(1 - \delta_{y,y_i}) \quad \forall y \in \mathcal{Y}, \forall i = 1 \ldots N$$

□
Chapter 5

Multiclass Classification with Simplex Least Squares

In this chapter, we show that the Simplex Least Squares (SLS) loss is classification calibrated, and derive explicit comparison inequalities and generalization bounds for the Regularized Least Squares algorithm. We present also in this chapter an experimental validation of the multiclass algorithms presented in this thesis on various benchmarks and datasets in Computer Vision.

5.1 Introduction

The use of least squares for binary classification has been investigated in a large body of work see for instance [RK04, BJM06, YRC07], and references there in. In the binary case, where classification amounts to a simple encoding/decoding using ±1 coding, the target function of least squares is given by

\[ f_\rho(x) = 2\rho_1(x) - 1, \]

and for any measurable function \( f \), we have the following comparison inequality:

\[ R(\text{sign}(f)) - R(b_\rho) \leq \sqrt{\mathcal{E}_V(f) - \mathcal{E}_V(f_\rho)} \]

i.e the link function \( \psi \) is given by \( \psi(t) = \sqrt{t}, t > 0 \). The comparison inequality for the square loss can be improved for a suitable class of probability distribution satisfying the so called Tsybakov noise condition [Tsy04],

\[ \mathbb{P}(\{x \in \mathcal{X}, |f_\rho(x)| \leq s\}) \leq B_\beta s^\beta, s \in [0, 1], \beta > 0, B_\beta > 0. \] (5.1)

Under this condition the probability of points mapped near 0 decreases polynomially. In this case the comparison inequality for the square loss is given by \( \psi(t) = c_\beta t^{\beta+1} \), see [BJM06, YRC07].
In this chapter we extend the analysis to the multiclass setting using the simplex coding, and show in Section 5.2 that the Simplex Least Squares (SLS) loss is indeed classification calibrated, and then derive comparison inequalities and introduce a similar condition to (5.1) for the multiclass setting. We then study in Section 5.3 the Regularized SLS, in vector valued RKHS, and derive generalization for the Regularized SLS algorithm. Algorithms and experimental validation of simplex coding based classification methods are finally presented in Sections 5.4 and 5.5.

5.2 SLS Calibration and Comparison Inequality

Recall that the SLS loss function is given by:

\[ V(y, w) = ||cy - w||^2. \]  

(5.2)

For \( p \in \Delta_q \), the inner risk \( \Psi_V(w, p) \) has the following form:

\[ \Psi_V(w, p) = \sum_{y \in Y} p_y ||cy - w||^2. \]  

(5.3)

In this section we show that, the minimizer \( w^* \) of \( \Psi_V(., p) \) is given by:

\[ w^* = \sum_{y \in Y} p_y cy. \]  

(5.4)

It follows that:

\[ \langle c_j, w^* \rangle = p_j - \frac{1}{q-1} \sum_{k \neq j} p_k = p_j - \frac{1}{q-1} (1 - p_j) = \frac{q}{q-1} p_j - \frac{1}{q-1}. \]  

(5.5)

Let

\[ f_\rho(x) = \sum_{y \in Y} \rho_y(x) cy, \forall x \in \mathcal{X} \]

from equation (5.5) it is easy to see now that,

\[ D(f_\rho(x)) = b_\rho(x), \quad b_\rho(x) = \arg \max_{j \in Y} \rho_j(x), \forall x \in \mathcal{X} (\ \text{up to sets of zero measure in } \rho_\mathcal{X}). \]

Hence we conclude that SLS is classification calibrated. The following theorem proves classification calibration of SLS and provides a comparison inequality relating for a measurable function \( f \) the excess misclassification risk to the excess SLS loss:

\[ \mathcal{E}_V(f) - \mathcal{E}_V(f_\rho) = \mathbb{E}(\Psi_V(f(x), \rho(x)) - \Psi_V(f_\rho(x), \rho(x))). \]

Theorem 4 (SLS Calibration and Comparison Inequality). Consider the SLS loss function
the following holds:

1. For $p \in \Delta_q^\Psi$, $\Psi_V(.,p)$ is minimized by $w^* = \sum_{j \in Y} p_j c_j$, we have $D(w^*) = y$, and SLS is classification calibrated.

2. Define $f_p(x) = \sum_{j \in Y} \rho_j(x) c_j \forall x \in \mathcal{X}$, then for any measurable function $f$ we have:

$$R(D(f)) - R(D(f_p)) \leq \sqrt{\frac{2(q-1)}{q} (\mathcal{E}_V(f) - \mathcal{E}_V(f_p))}.$$  \hspace{1cm} (5.6)

**Proof of Theorem 4.** The proof of Theorem 4 is given in Section 5.6. \hfill \Box

Note that unlike DC-SVM, where the comparison inequality involved the excess DC-SVM loss in a linear way ($\psi(t) = Ct$), the comparison inequality for SLS involves the square root of the excess Least Squares loss ($\psi(t) = C\sqrt{t}$).

In the next section we ask the question if this exponent can be enhanced for SLS under certain low noise condition on the labels.

### 5.2.1 Improved Exponents under Noise Condition

The above theorem can be improved for certain classes of distribution. Toward this end we introduce the following notion of misclassification noise that generalizes Tsybakov’s noise condition [Tsy04].

**Definition 8 (Generalized Tsybakov Noise).** Fix $\beta > 0$, we say that the distribution $\rho$ satisfy the classification noise condition with parameter $B_\beta > 0$, if

$$\rho_x \left( \left\{ x \in \mathcal{X} \mid 0 \leq \min_{j \neq D(f_p(x))} \frac{q-1}{q} \langle cD(f_p(x)) - c_j, f_p(x) \rangle \leq s \right\} \right) \leq B_\beta s^\beta, \hspace{1cm} (5.7)$$

where $s \in [0, 1]$.

If a distribution $\rho$ is characterized by a very large $\beta$, then, for each $x \in \mathcal{X}$, $f_\rho(x)$ is arbitrarily close to one of the coding vectors.

For $q = 2$, the above condition reduces to the binary Tsybakov noise. Indeed, let $c_1 = 1$, and $c_2 = -1$, if $f_\rho(x) > 0$, $\frac{1}{2} (c_1 - c_2) f_\rho(x) = f_\rho(x)$, and if $f_\rho(x) < 0$, $\frac{1}{2} (c_2 - c_1) f_\rho(x) = -f_\rho(x)$.

We have the following result:

**Theorem 5 (Low Noise Comparison Inequality).** For each any measurable function $f$, if (5.7) holds, then we have the following inequality,

$$R(D(f)) - R(D(f_p)) \leq K \left( \frac{2(q-1)}{q} (\mathcal{E}(f) - \mathcal{E}(f_p)) \right)^{\frac{q+1}{q^2+2}}, \hspace{1cm} (5.8)$$

for a constant $K = (2\sqrt{B_\beta + 1})^{-\frac{2(q+2)}{q^2+2}}$. 


Proof of Theorem 5. The proof of Theorem 5 is given in Section 5.6.

Note that the exponent \( \frac{\beta+1}{\beta+2} \), is an increasing function for \( \beta > 0 \), and tends to 1 at infinity, hence a large \( \beta \), ensures a linear exponent in the comparison inequality for SLS, and approaches the exponent of DC-SVM.

5.3 Regularized SLS in \( vv \)-RKHS

5.3.1 Regularized SLS

In Chapter 4 we reviewed the regularization in \( vv \)-RKHS. Let \( \mathcal{H}_\Gamma \), be a vector valued RKHS induced by the kernel \( \Gamma \). Similarly to DC-SVM, we have the following expected risk minimization:

\[
\min_{f \in \mathcal{H}_\Gamma} \mathbb{E}(||c_y - f(x)||^2) + \lambda ||f||^2_{\mathcal{H}_\Gamma},
\]

let \( f^\lambda \) be its minimizer. Given a training set \( S = \{(x_i, y_i), x_i \in \mathcal{X}, y_i \in \mathcal{Y}, i = 1 \ldots N\} \), the regularized empirical risk is:

\[
\min_{f \in \mathcal{H}_\Gamma} \frac{1}{N} \sum_{i=1}^{N} ||c_{y_i} - f(x_i)||^2 + \lambda ||f||^2_{\mathcal{H}_\Gamma},
\]

let \( f^\lambda_N \) be its minimizer. The following theorem assesses the convergence of \( R(D(f^\lambda_N)) \) to \( R(D(f_\rho)) \).

5.3.2 Generalization Bounds

Theorem 6 (Generalization Bound). Assume \( \mathcal{H}_\Gamma \) is universal \( vv \)-RKHS, such that:

\[
\mathcal{E}_V(f_\rho) = \inf_{f \in \mathcal{H}_\Gamma} \mathcal{E}_V(f).
\]

Let \( \Gamma_{\max} = \sup_{x \in \mathcal{X}} ||\Gamma(x, x')|| \), and \( A(\lambda) = \inf_{f \in \mathcal{H}_\Gamma} \mathcal{E}_V(f) + \lambda ||f||^2_{\mathcal{H}_\Gamma} - \inf_{f \in \mathcal{H}_\Gamma} \mathcal{E}_V(f) \). The following holds for SLS,

1. Let \( \delta > 0 \), there exists \( C(\delta) > 0 \) (\( C(\delta) \) is independent of \( q \))

\[
\mathcal{E}_V(f^\lambda_N) - \inf_{f \in \mathcal{H}_\Gamma} \mathcal{E}_V(f) \leq C(\delta) \left( A(\lambda) + \sqrt{\frac{\Gamma_{\max}}{\lambda N}} \right),
\]

with probability \( 1 - \delta \).
2. The following holds with probability \(1 - \delta\),

\[
R(D(\hat{f}^N_\lambda)) - R(b_\rho) \leq \sqrt{\frac{2(q - 1)}{q}} \left(C(\delta) \left( A(\lambda) + \sqrt{\frac{\Gamma_{\max}}{\lambda N}} \right) \right)^{\frac{1}{2}}, \tag{5.11}
\]

3. If the distribution \(\rho\) satisfies the low noise condition given in Definition 5.7 with constant \(K > 0\), and exponent \(\beta > 0\) then:

\[
R(D(\hat{f}^N_\lambda)) - R(b_\rho) \leq K \left(\frac{2(q - 1)}{q} C(\delta) \left( A(\lambda) + \sqrt{\frac{\Gamma_{\max}}{\lambda N}} \right) \right)^{\frac{\beta + 1}{\beta + 2}}. \tag{5.12}
\]

Proof. 1) The proof is along the line of proofs in [CV07] where sharper bounds are given using more advanced techniques. 2) and 3) follow from combining (5.10), and Theorems 4, and 5.

Note that for both DC-SVM and SLS our bounds are not sharp, and can be improved using more advanced techniques such as localized Rademacher complexities [BBM02], or bounded variance assumptions on the risks of the estimators [CV07, SC08]. These bounds, while not sharp give an indication on the importance of the exponent in the comparison inequality and how it transfers to the sample complexity of the algorithm.

5.4 Algorithms

Let \(\Gamma\) be a separable kernel:

\[
\Gamma(x, x') = k(x, x')I_{(q-1)},
\]

where \(k\) is a scalar kernel. Appealing to the representer theorem we know that:

\[
f^N_\lambda(x) = \sum_{i=1}^{N} k(x, x_i)a_i, \quad a_i \in \mathbb{R}^{q-1}, \quad i = 1 \ldots N,
\]

and the norm of \(f^N_\lambda\) has the following form,

\[
||f||_{\mathcal{H}_\Gamma}^2 = \sum_{i=1}^{N} \sum_{j=1}^{N} k(x_i, x_j) \langle a_i, a_j \rangle.
\]

Let \(K \in \mathbb{R}^{N \times N}\) be the kernel matrix, where \(K_{i,j} = k(x_i, x_j), i, j \in \{1 \ldots N\}\). Let \(Y \in \mathbb{R}^{N \times (q-1)}\) be the label matrix, where the \(i\)-th row of \(Y\), \(Y_{i, \cdot} = c_{y_i}, i = 1 \ldots N\). Define \(A \in \mathbb{R}^{N \times (q-1)}\), such that \(A = [a_1, \ldots, a_N]^T\). Problem (5.9), can be written with matrix
notation, in the following way:
\[
\min_{A \in \mathbb{R}^{N \times (q-1)}} \frac{1}{N} \|Y - KA\|_F^2 + \lambda \text{Tr}(A^T KA). \tag{5.13}
\]

It is easy to see that in this case the coefficients must satisfy
\[
(K + \lambda NI_N)A = Y. \tag{5.14}
\]

We turn now to the linear case where we have an explicit feature map \( \Phi : \mathcal{X} \to \mathbb{R}^p \). Let \( X \in \mathbb{R}^{N \times p}, X = [\Phi(x_1), \ldots, \Phi(x_N)]^\top \), in that case it easy to see that we are looking for a linear function:
\[
f(x) = W^\top \Phi(x), \quad W \in \mathbb{R}^{p \times (q-1)},
\]
and problem (5.9) is equivalent to:
\[
\min_{W \in \mathbb{R}^{p \times (q-1)}} \frac{1}{N} \|Y - XW\|_F^2 + \lambda \|W\|_F^2. \tag{5.15}
\]

It is easy to see that solution of (5.15) satisfies:
\[
(X^\top X + \lambda NI_{pq})W = X^\top Y. \tag{5.16}
\]

Equations (5.14) and (5.16), can be solved efficiently using elementary linear algebra, we refer the reader to [TMSR13] for a succinct review.

### 5.4.1 Regularization Path and Model Selection

Interestingly, the classical results from [Wah90] can be extended to show that the value \( f_{S_i}(x_i) \), obtained computing the solution \( f_{S_i} \) removing the \( i \)-th point from the training set (the leave one out solution), can be computed in closed form.

Let \( f_{\text{loo}}^\lambda \in \mathbb{R}^{N \times (q-1)} \),
\[
f_{\text{loo}}^\lambda = (f_{S_1}^\lambda(x_1), \ldots, f_{S_N}^\lambda(x_N)).
\]

Let \( K(\lambda) = (K + \lambda nI)^{-1} \) and \( C(\lambda) = K(\lambda)Y \). Define \( M(\lambda) \in \mathbb{R}^{N \times (q-1)} \), such that:
\[
M(\lambda)_{ij} = 1/K(\lambda)_{ii}, \quad \forall \ j = 1 \ldots q - 1.
\]

One can show that \( f_{\text{loo}}^\lambda = Y - C(\lambda) \odot M(\lambda) \), where \( \odot \) is the Hadamard product. Then, the leave-one-out error
\[
\frac{1}{N} \sum_{i=1}^{N} \mathbb{I}_{y \neq D(f_{S_i}(x))}(x_i, y_i),
\]

can be minimized at essentially no extra cost by precomputing the eigen decomposition of \( K \) (or \( X^\top X \)).

### 5.4.2 Online/Incremental Optimization

The regularized estimators induced by the simplex loss functions can be computed by mean of online/incremental first order (sub) gradient methods. Indeed, when considering finite dimensional feature maps, these strategies offer computationally feasible solutions to train
estimators for large datasets where neither a $p$ by $p$ nor an $N$ by $N$ matrix fit in memory. Following [SSSS07] we can alternate a step of stochastic descent on a data point: $W_{\text{tmp}} = (1 - \eta_i \lambda) W_i - \eta_i \partial(V(y_i, W_i^T \Phi(x_i)))$ and a projection on the Frobenius ball $W_i = \min(1, \frac{1}{\sqrt{\lambda ||W_{\text{tmp}}||_F}}) W_{\text{tmp}}$. The algorithm depends on the used loss function through the computation of the (point-wise) subgradient $\partial(V)$. The latter can be easily computed for all the loss functions previously discussed.

For the SLS loss we have $\partial(V(y_i, f_W(x_i))) = 2 \Phi(x_i) (c_{y_i} - W^T \Phi(x_i))^T$. For the DC-SVM loss we have

$$\partial(V(y_i, f_W(x_i))) = \Phi(x_i) \left( \sum_{k \in I_i} c_k \right)^T$$

where $I_i = \{y \neq y_i | \langle c_y, W^T \Phi(x_i) \rangle > -\frac{1}{q-1} \}$. For the HS-SVM loss we have:

$$\partial(V(y, f_W(x_i))) = -\Phi(x_i)c_{y_i}^T \text{ if } \langle c_{y_i}, W^T \Phi(x_i) \rangle < 1 \text{ and } 0 \text{ else }.$$ 

For AC-SVM, $\partial(V(y, f_W(x_i))) = -\Phi(x_i) \left( \sum_{k \in I} (c_{y_i} - c_k) \right)^T$, $I = \{y \neq y_i, \langle c_y, c_{y_i}, W^T \Phi(x_i) \rangle < \frac{q}{q-1} \}$.

5.5 Numerical Validation for Classification via Simplex Coding

5.5.1 Comparison of Computational Complexity

The cost of solving S-RLS for fixed $\lambda$ is in the worst case $O(N^3)$ (for example via Choleski decomposition). If we are interested into computing the regularization path for $R$ regularization parameter values, then as noted in [RK04] it might be convenient to perform an eigendecomposition of the kernel matrix rather than solving the systems $R$ times. For explicit feature maps the cost is $O(Np^2)$, so that the cost of computing the regularization path for simplex RLS algorithm is $O(\min(N^3, Np^2))$ and hence independent of $q$. One can contrast this complexity with the one of a naïve One Versus all (OVA) approach that would lead to a $O(RN^3q)$ complexity. Simplex SVMs can be solved using solvers available for binary SVMs that are considered to have complexity $O(N\gamma)$ with $\gamma \in \{2, 3\}$ (actually the complexity scales with the number of support vectors). For DC-SVM, though, we have $Nq$ rather than $N$ unknowns and the complexity is $O(Nq\gamma)$. Note that unlike for S-RLS, there is no straightforward way to compute the regularization path and the leave one out error for any of the simplex SVMs. The online algorithms induced by the different simplex loss functions are essentially the same, in particular each iteration depends linearly on the number of classes.
5.5.2 Numerical Results

We conduct several experiments to evaluate the performance of our batch and online algorithms, on 5 UCI datasets, as well as on Caltech101 and on a face verification dataset Pubfig83. See table below for details:

<table>
<thead>
<tr>
<th></th>
<th>ntrain</th>
<th>p</th>
<th>q</th>
<th>ntest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Landsat</td>
<td>4435</td>
<td>36</td>
<td>6</td>
<td>2000</td>
</tr>
<tr>
<td>Optdigit</td>
<td>3823</td>
<td>64</td>
<td>10</td>
<td>1797</td>
</tr>
<tr>
<td>Pendigit</td>
<td>7494</td>
<td>16</td>
<td>121</td>
<td>3498</td>
</tr>
<tr>
<td>Letter</td>
<td>10000</td>
<td>16</td>
<td>26</td>
<td>10000</td>
</tr>
<tr>
<td>Isolet</td>
<td>6238</td>
<td>617</td>
<td>26</td>
<td>1559</td>
</tr>
<tr>
<td>Pubfig83</td>
<td>7470</td>
<td>25600</td>
<td>83</td>
<td>830</td>
</tr>
<tr>
<td>Ctech101</td>
<td>3060</td>
<td>8192</td>
<td>102</td>
<td>6084</td>
</tr>
</tbody>
</table>

We compare the performance of our algorithms to on versus all SVM (libsvm), as well as the simplex based boosting [SN11]. For UCI datasets we use the raw features, on Caltech101 we used HMAX features, and on Pubfig83 we use the feature maps from [PSZC11]. In all cases the parameter selection is based either on a hold out (ho) (80% training - 20% validation) or a leave one out error (loo). For the model Selection of $\lambda$ in S-LS, 100 values are chosen in the range $[\lambda_{\min}, \lambda_{\max}]$, (where $\lambda_{\min}$ and $\lambda_{\max}$ correspond to the smallest and biggest eigenvalues of $K$). In the case of a Gaussian kernel (RBF) we use a heuristic that sets the width of the gaussian $\sigma$ to the 25-th percentile of pairwise distances between distinct points in the training set. In Table 5.1 we collect the resulting classification accuracies:

<table>
<thead>
<tr>
<th></th>
<th>Landsat</th>
<th>Optdigit</th>
<th>Pendigit</th>
<th>Letter</th>
<th>Isolet</th>
<th>Ctech</th>
<th>Pubfig83</th>
</tr>
</thead>
<tbody>
<tr>
<td>DC-SVM Online (ho)</td>
<td>65.15%</td>
<td>89.57%</td>
<td>81.62%</td>
<td>52.82%</td>
<td>88.58%</td>
<td>63.33%</td>
<td>84.70%</td>
</tr>
<tr>
<td>HS-SVM Online (ho)</td>
<td>75.43%</td>
<td>85.58%</td>
<td>72.54%</td>
<td>38.40%</td>
<td>77.65%</td>
<td>45%</td>
<td>49.76%</td>
</tr>
<tr>
<td>SLS Online (ho)</td>
<td>63.62%</td>
<td>91.68%</td>
<td>81.39%</td>
<td>54.29%</td>
<td>92.62%</td>
<td>58.39%</td>
<td>83.61%</td>
</tr>
<tr>
<td>SLS Batch (loo)</td>
<td>65.88%</td>
<td>91.90%</td>
<td>80.69%</td>
<td>54.96%</td>
<td>92.55%</td>
<td>66.35%</td>
<td>86.63%</td>
</tr>
<tr>
<td>SLS rbf Batch (loo)</td>
<td>90.15%</td>
<td>97.09%</td>
<td>98.17%</td>
<td>96.48%</td>
<td>97.05%</td>
<td>69.38%</td>
<td>86.75%</td>
</tr>
<tr>
<td>SVM batch ova (ho)</td>
<td>72.81%</td>
<td>92.13%</td>
<td>86.93%</td>
<td>62.78%</td>
<td>90.59%</td>
<td>70.13%</td>
<td>85.97%</td>
</tr>
<tr>
<td>SVM rbf batch ova (ho)</td>
<td>95.33%</td>
<td>98.07%</td>
<td>98.88%</td>
<td>97.12%</td>
<td>96.99%</td>
<td>51.77%</td>
<td>85.60%</td>
</tr>
<tr>
<td>Simplex boosting [SN11]</td>
<td>86.65%</td>
<td>92.82%</td>
<td>92.94%</td>
<td>59.65%</td>
<td>91.02%</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 5.1: Accuracies of our algorithms on several datasets.

As suggested by the theory, the consistent methods DC-SVM and S-LS have a big advantage over HS-SVM. Batch methods are overall superior to online methods, with online DC-SVM achieving the best results. More generally, we see that rbf SLS has the best performance among the simplex methods including the simplex boosting [SN11]. When compared to One Versus All SVM-rbf, we see that SLS rbf achieves essentially the same performance.
5.6 Proofs of Theorem 4 and Theorem 5

Proof of Theorem 4. 1) The proof is straightforward, setting derivative to zero.
2) As \( f_p(x) = \sum_{y \in \mathcal{Y}} c_y \rho_y(x) \) we can write,
\[
\langle f_p(x), c_y \rangle = \sum_{y' \in \mathcal{Y}} \rho_{y'}(x) \langle c_{y'}, c_y \rangle = \rho_y(x) - \frac{1}{q-1} (1 - \rho_y(x)) = \frac{q \rho_y(x) - 1}{q-1}
\]
so that
\[
\rho_y(x) = \frac{q-1}{q} \langle f_p(x), c_y \rangle + \frac{1}{q}, \quad \forall \ y \in \mathcal{Y}.
\]
(5.17)

Fisher consistency easily follows from the definition of the decoding map \( D \). We next derive the comparison inequality (5.6). We begin noting that by definition,
\[
R(D(f)) = \int_{\mathcal{X}} d\rho_x(x) \sum_{y \in \mathcal{Y}} \mathbb{I}_{y \neq D(f(x))}(x, y) = \int_{\mathcal{X}} d\rho_x(x) \sum_{y \neq D(f(x))} \rho_y(x) = \int_{\mathcal{X}} d\rho_x(x)(1 - \rho_{D(f(x))}(x))
\]
so that,
\[
R(D(f)) - R(D(f_p)) = \int_{\mathcal{X}_f} \left( \rho_{D(f_p(x))}(x) - \rho_{D(f(x))}(x) \right) d\rho_x(x) = \int_{\mathcal{X}_f} \left( \rho_{D(f_p(x))}(x) - \rho_{D(f(x))}(x) \right) d\rho_x(x)
\]
(5.18)

where \( \mathcal{X}_f = \{ x \in \mathcal{X} \mid D(f(x)) \neq D(f_p(x)) \} \). Moreover using equation (5.17) we can write,
\[
R(D(f)) - R(D(f_p)) = \frac{q-1}{q} \int_{\mathcal{X}_f} \langle c_{D(f_p(x))} - c_{D(f(x))}, f_p(x) \rangle d\rho_x(x)
\]
\[
= \frac{q-1}{q} \int_{\mathcal{X}_f} \langle c_{D(f_p(x))} - c_{D(f(x))}, f_p(x) - f(x) \rangle d\rho_x(x)
\]
\[
+ \frac{q-1}{q} \int_{\mathcal{X}_f} \langle c_{D(f_p(x))} - c_{D(f(x))}, f(x) \rangle d\rho_x(x).
\]
(5.19)

The last term in the above expression can be shown to be negative since
\[
\langle c_{D(f_p(x))}, f(x) \rangle \leq \langle c_{D(f(x))}, f(x) \rangle = \max_{y \in \mathcal{Y}} \langle c_y, f(x) \rangle, \quad \forall \ x \in \mathcal{X}_f,
\]
(5.20)
by definition of $D$. Then using Jensen and Cauchy-Schwarz and inequalities we have,

$$
(R(D(f)) - R(D(f_\rho)))^2 \leq \frac{(q-1)^2}{q^2} \int_{X_f} \left( \langle c_{D(f_\rho(x))} - c_{D(f(x))}, f_\rho(x) - f(x) \rangle \right)^2 d\rho_X(x)
$$

$$
\leq \frac{(q-1)^2}{q^2} \int_{X_f} \|c_{D(f_\rho(x))} - c_{D(f(x))}\|^2 \|f(x) - f_\rho(x)\|^2 d\rho_X(x)
$$

$$
\leq \frac{(q-1)^2}{q^2} \frac{2q}{q-1} \int_{X_f} \|f(x) - f_\rho(x)\|^2 d\rho_X(x)
$$

$$
\leq \frac{2(q-1)}{q} \int \|f(x) - f_\rho(x)\|^2 d\rho_X(x), \tag{5.21}
$$

where we used the fact that $\|c_y - c'_y\|^2 = \frac{2q}{q-1}$, for $y \neq y'$.

On the other hand note that:

$$
\mathcal{E}_V(f) = \mathbb{E}[\|Y - f(X)\|^2] = \mathbb{E}[\|Y - f_\rho\| + (f_\rho(X) - f(X))|^2] = \|f - f_\rho\|_p^2 + \mathcal{E}_V(f_\rho), \tag{5.22}
$$

since

$$
-2 \mathbb{E}[\langle Y - f_\rho, f_\rho - f \rangle_{\rho}] = \int d\rho_X(x) \left( \sum_{y \in Y} c_y \rho_Y(x) - f_\rho(x), f_\rho(x) - f(x) \right) = 0.
$$

The result follows plugging (5.22) in the above expression in (5.21) and taking square roots. \hfill \Box

**Proof of Theorem 5.** We start with the following lemma:

**Lemma 6.** The generalised Tsybakov condition is equivalent to that for any measurable function $f$:

$$
\rho(X_f) \leq C_\alpha (R(D(f)) - R(D(f_\rho)))^\alpha \tag{5.23}
$$

where $\alpha = \frac{q}{\beta + 1} < 1$ and $C_\alpha = B_\beta + 1 > 1$.

Let $0 < t < 1$ if $t \leq m_\rho(x)$ we have $tm_\rho(x) \leq m_\rho^2(x)$ and therefore $m_\rho(x) \leq \frac{m_\rho^2(x)}{t}$.

$$
R(D(f)) - R(D(f_\rho)) = \int_{X_f} m_\rho(x) d\rho_X(x)
$$

$$
= \int_{X_f} \rho(x) 1_{m_\rho(x) \leq t} d\rho_X(x) + \int_{X_f} \rho(x) 1_{m_\rho(x) > t} d\rho_X(x)
$$

$$
\leq t \rho(X_f) + \frac{1}{t} \int_{X_f} \rho(x)^2 d\rho_X(x)
$$

$$
\leq tC_\alpha ((R(D(f)) - R(D(f_\rho)))^\alpha + \frac{1}{t} \frac{2(q-1)}{q} (\mathcal{E}_V(f) - \mathcal{E}_V(f_\rho))
$$
In the last inequality we used lemma 6 and the fact:

\[ m_p(x) = \frac{q - 1}{q} \langle c_{D(f_p(x))} - c_{D(f(x))}, f_p(x) \rangle \leq \frac{q - 1}{q} \langle c_{D(f_p(x))} - c_{D(f(x))}, f_p(x) - f(x) \rangle. \]

Squaring both sides of the inequality and using Cauchy Schwarz:

\[ m_p(x)^2 \leq \left( \frac{q - 1}{q} \right)^2 \frac{q}{q - 1} ||f(x) - f_p(x)||^2, \]

and

\[
\int_{\mathcal{X}_f} m_p^2(x) d\rho_X(x) \leq \frac{2(q - 1)}{q} \int_{\mathcal{X}_f} ||f(x) - f_p(x)||^2 d\rho_X(x)
\leq \frac{2(q - 1)}{q} ||f - f_p||_p^2 = \frac{2(q - 1)}{q} (\mathcal{E}_f - \mathcal{E}_p)
\]

Minimizing the right hand side of the above inequality over \( t \), we get the result (5.8). \( \Box \)

**Proof of lemma 6.** Let \( m_p(x) = \frac{q - 1}{q} \langle c_{D(f_p(x))} - c_{D(f(x))}, f_p(x) \rangle \)

\[
R(D(f)) - R(D(f_p)) = \int_{\mathcal{X}_f} m_p(x) d\rho_X(x) \geq \int_{\mathcal{X}_f} m_p(x) \mathbb{1}_{m_p(x) \geq t} d\rho_X(x),
\]

\[
\geq t \left( \int_{\mathcal{X}} \mathbb{1}_{m_p(x) \geq t} d\rho_X(x) - \int_{\mathcal{X}/\mathcal{X}_f} \mathbb{1}_{m_p(x) \geq t} d\rho_X(x) \right)
\geq t(1 - \mathbb{P}\{x \in \mathcal{X}, m_p(x) \leq t\} - \mathbb{P}\{x \in \mathcal{X}/\mathcal{X}_f\})
\geq t(1 - B_\beta t^\beta - \rho_X(\mathcal{X}/\mathcal{X}_f))
\]

Now taking the minimum of the above bound with respect to \( t \), we get \( t^* = \left( \frac{1}{B_\beta + 1} \rho_X(\mathcal{X}_f) \right)^{\frac{1}{\beta}} \).

Finally plugging \( t^* \) in the bound we get,

\[ \rho_X(\mathcal{X}_f) \leq C_\alpha (R(D(f)) - R(D(f_p)))^\alpha \quad (5.24) \]

where \( \alpha = \frac{\beta}{\beta + 1} < 1 \) and \( c_\alpha = B_\beta + 1 > 1 \). \( \Box \)
Chapter 6

Deep Multimodal Learning

While the first chapters of this thesis focused on multiple outputs prediction in a multiclass setting, we turn in this chapter to the problem of multiple inputs/multiple outputs prediction. More specifically in this chapter, we present several methods in deep multimodal learning for fusing speech and visual modalities in Audio-Visual Automatic Speech Recognition (AV-ASR).

6.1 Introduction

Human speech perception is not only about hearing but also about seeing: our brain integrates the waveforms representing the speech information as well as the lips poses and motions, often called visemes, which carry important visual information about what is being said. This has been demonstrated by the so called McGurk effect [MM76], which shows that a voicing of $ba$ and a mouthing of $ga$ is perceived as being $da$. In the presence of noise and multiple speakers (cocktail party effect), humans rely on lip reading in order to enhance speech recognition [CHLN08]. The visual information is also important in a clean speech scenario as it helps in disambiguating voices with similar acoustics [Sum92].

In Audio-Visual Automatic Speech Recognition (AV-ASR), both audio recordings and videos of the person talking are available at training time. It is challenging to build models that integrates both visual and audio information, and that enhance the recognition performance of the overall system. While most previous works in AV-ASR focused on enhancing the performance in the noisy case [PNLM04, NKK+11], where the visual information can be crucial, we focus in this chapter on showing that the visual information is indeed helpful even in the clean speech scenario.

Multimodal learning consists of fusing and relating information coming from different sources, hence AV-ASR is an important multimodal problem. Finding correlations between different modalities, and modeling their interactions, has been addressed in various learning frameworks and has been applied to AV-ASR [Gea09, LS06, MHD96, PKPM07, PKPM09, PKPM06]. Deep Neural Networks (DNN) have shown impressive performance in both audio
and visual classification tasks, which is why we restrict ourselves to the deep multimodal learning framework [YGS89, NKK+11, SS12, SCMN13, AALB13].

In this chapter, we propose methods in deep learning to fuse modalities, and validate them on the IBM AV-ASR Large Vocabulary Studio Dataset (Section 6.2). First we consider the training of two networks on the audio and the visual modality separately. Then, considering the last layer of each network as a better feature space, and concatenating them, we train a classifier on that joint representation, and obtain gains in Phone Error Rates (PER), with respect to an audio-only trained network. We then propose a new bilinear network that accounts for correlations between modalities and allows for joint training of the two networks, we show that a committee of such bilinear networks, fused at the level of posteriors, achieves a better PER in a clean speech scenario.

The chapter is organized as follows. In Section 6.2 we present the IBM AV-ASR large vocabulary studio dataset, our feature extraction pipeline for the audio and the visual channels. Next, in Section 6.3, we present results for the fusion of networks separately trained on each modality. In Section 6.4 we introduce the bilinear DNN that allows for a joint training and captures correlations between the two modalities, and derive its back-propagation algorithm in Section 6.5. Finally we present posterior combination of bimodal and bilinear bimodal DNNs in Section 6.6.

6.2 Audio-Visual Data Set & Feature Extraction

In this Section we present the IBM AV-ASR Large Vocabulary Studio dataset, and our feature extraction pipeline.

6.2.1 IBM AV-ASR Large Vocabulary Studio Dataset

The IBM AV-ASR Large Vocabulary Studio Dataset consists of 40 hours of audio-visual recordings from 262 speakers. These were carried out in clean, studio conditions. The audio is sampled at 16 KHz along with the video frame rate of 30 frames per second at 704 x 480 resolution. The vocabulary size in these recordings is 10,400 words. This data set was divided into a test set of 2 hours of audio+video from 22 speakers, with the rest used for training.

6.2.2 Feature Extraction

For the audio channel we extract 24 MFCC coefficients at 100 frames per second. Nine consecutive frames of MFCC coefficients are stacked and projected to 40 dimensions using an LDA matrix. Input to the audio neural network is formed by concatenating ±4 LDA frames to the central frame of interest, resulting in an audio feature vector of dimension 360.

For the visual channel we start by detecting the face in the image using the openCV implementation of the Viola-Jones algorithm. We then do a mouth carving by an openCV
mouth detection model. Both these utilize the ENCARA2 model as described in [MDHL11]. In order to get an invariant representation to small distortions and scales we then extract level 1 and level 2 scattering coefficients [BM13] on the 64 x 64 mouth region of interest and then reduce their dimension to 60 using LDA (Linear discriminant Analysis). In order to match the audio frame rate we replicate video frames according to audio and video time stamps. We also add ±4 context frames to the central frame of interest, and obtain finally a visual feature vector of dimension 540.

![Fig 6-1: Low Level Features Extraction for the Audio and the Visual Channels.](image)

**6.2.3 Context-dependent Phoneme Targets**

Each audio+video frame is labeled with one of 1328 targets that represent context dependent phonemes. 42 phones in phonetic context of ±2 are clustered using decision trees down to 1328 classes. We measure classification error rate at the level of these 1328 classes, this is referred to as phone error rate (PER).

**6.3 Uni-modal DNNs & Feature Fusion**

In the supervised multimodal scenario, we are given a training set $S$ of $N$ labeled examples, and $C$ classes:

$$S = \{(x^1_i, x^2_i, y_i), i = 1...N\}, \quad y_i \in \mathcal{Y} = \{1...C\},$$

where $x^1_i, x^2_i$ correspond to the first and the second modality feature vectors, respectively. We note $t_i = e_{y}$, the classification targets, where $\{e_y\}_{y \in \mathcal{Y}}$ is the canonical basis in $\mathbb{R}^C$. Let $\rho(y|x^1, x^2)$ be the posterior probability of being in class $y$ given the two modalities $x^1$ and $x^2$. In a classification task, we would like to find the model that maximizes the cross-entropy.
The first multimodal modeling approach we study is to train two separate networks $DNN_a$ and $DNN_v$, on the audio and the visual features, respectively. The networks are optimized under the cross-entropy objective (6.1) using the stochastic gradient descent. We formed a joint Audio-Visual feature representation by concatenating the outputs of final hidden layers of these two networks, as shown in Figure 6.3. This feature space is then kept fixed while a deep or a shallow (softmax only) network is trained in this fused space up to the targets. To keep the feature space dimension manageable, we configure the individual audio and video networks to have a low dimensional final hidden layer.

We consider for $DNN_a$ and $DNN_v$ the following architecture

\[ dim/1024/1024/1024/1024/1024/200/1328, \]

where $dim = 360$ for $DNN_a$ and $dim = 540$ for $DNN_v$. The fused feature space dimension is 400.

While $DNN_a$ achieves a PER of 41.25%, $DNN_v$ alone achieves a PER of 69.36%, showing that the visual information alone carries some information but that is not enough in itself to get a low error rate. A deep network built in the fused feature space results in a PER of 35.77% while a softmax layer only in this feature space yields PER of 35.83%. This substantial PER gain from joint audio-visual representation, even in clean audio conditions, demonstrates the value of visual information for the phoneme classification task. Interes-
ingly, the deep and the shallow fusion are roughly on par in terms of PER. Results are summarized in the following table:

<table>
<thead>
<tr>
<th></th>
<th>PER</th>
<th>Cross-Entropy</th>
</tr>
</thead>
<tbody>
<tr>
<td>$DNN_a$ (Audio Alone)</td>
<td>41.25%</td>
<td>1.53948848</td>
</tr>
<tr>
<td>$DNN_v$ (Visual Alone)</td>
<td>69.36%</td>
<td>3.24791566</td>
</tr>
<tr>
<td>Bimodal (DNN Fusion)</td>
<td><strong>35.77%</strong></td>
<td><strong>1.31047744</strong></td>
</tr>
<tr>
<td>Bimodal (SoftMax Fusion)</td>
<td><strong>35.83%</strong></td>
<td><strong>1.31077926</strong></td>
</tr>
</tbody>
</table>

Table 6.1: Empirical Evaluation on the AV-ASR Studio dataset.

### 6.4 Bilinear Deep Neural Network

In the previous section the training was done separately on the two modalities, in this section we address the joint training problem, and introduce the bilinear bimodal DNN.

For a DNN, we note by $\sigma$ the non linearity function (sigmoid in this chapter), $v_{\ell}$ the input of a unit, and $h_{\ell}$ the output of a unit in a layer $\ell$. For a layer $\ell$ we note the dimension of an input $v_{\ell}$ by $K_{\ell}$. As shown in Figure 6.3, we consider two DNNs, one for each modality that we fuse at the level of the decision function. For simplicity of the exposure, we assume the same number of layers $L (L_1 = L_2 = L)$. For the intermediate layers, we have the standard separate networks:

$$h_{\ell}^j = \sigma(W_{\ell}^j v_{\ell}^j + b_{\ell}^j), \quad v_{\ell}^0 = x_{\ell}^j,$$

$$W_{\ell}^j \in \mathbb{R}^{K_{\ell}^j \times K_{\ell+1}^j}, b_{\ell}^j \in \mathbb{R}^{K_{\ell+1}^j}, \ell = 1 \ldots L - 1, j \in \{1, 2\}.$$

The fusion happens at the last hidden layer, where the posteriors capture the correlation between the intermediate non-linear features of the two modalities produced by the DNN layers, through a bilinear term. Let $v_{L}^1 = h_{L - 1}^1$, $v_{L}^2 = h_{L - 1}^2$, the posteriors have the following form:

$$\rho(y|x^1, x^2) = \frac{\exp \left( v_{L}^{1,\top} W y v_{L}^2 + V_y^\top \left( \begin{array}{c} v_{L}^1 \\ v_{L}^2 \end{array} \right) + b_y \right)}{Z},$$

(6.2)

where $Z = \sum_{y' \in Y} \exp \left( v_{L}^{1,\top} W y' v_{L}^2 + V_y^\top \left( \begin{array}{c} v_{L}^1 \\ v_{L}^2 \end{array} \right) + b_{y'} \right), W y \in \mathbb{R}^{K_L^1 \times K_L^2}, V_y = [V_y^1, V_y^2] \in \mathbb{R}^{(K_L^1 + K_L^2)}, b_y \in \mathbb{R}$ and $y \in \{1 \ldots C\}$.

### 6.4.1 Factored Bilinear Softmax

As the number of classes increases, the bilinear model becomes cumbersome computationally, and we need large training sets to get better estimates of the parameters. In order to decrease the computational complexity of the model, we propose the use of a factorization of the bilinear term, that is similar to the one in [MZHP10], but is motivated in our case by
Canonical Correlation Analysis (CCA) [HSST04]:

\[ W^y = U^1 \text{diag}(w_y) U^2, y = 1 \ldots C, \] (6.3)

where \( U^1 \in \mathbb{R}^{K^1 \times F} \), \( U^2 \in \mathbb{R}^{K^2 \times F} \), \( w_y \in \mathbb{R}^F \), and \( \text{diag}(w_y) \) is a diagonal matrix with \( w_y \) on its diagonal. For numerical stability we consider \( ||U^j||_F \leq \lambda, j \in \{1, 2\} \), where \( \lambda \) is a regularization parameter. We note by \( F \) the dimension of the fused space, which is typically smaller than \( K^1 \) and \( K^2 \). Considering the factorization in (6.3) and maximizing the cross-entropy in the bilinear model (6.2), we have:

\[
\log \rho(y|x^1, x^2) = \text{Tr}(U^1 v^1_L v^2_L^T U^2 \text{diag}(w_y)) + \langle V^1_y, v^1_L \rangle + \langle V^2_y, v^2_L \rangle + b_y - \log(Z).
\]

For fixed weights \( w_y \), learning \((U^1, U^2)\) corresponds to a class specific weighted CCA-like learning where we are looking for projections that maximize alignment between the intermediate features of the two modalities, in a discriminative way. Deep CCA of [AALB13] shares similarities with this model.

On the other hand, for fixed \((U^1, U^2)\), we can rewrite the log-posteriors in the following way:

\[
\log(\rho(y|x^1, x^2)) = \langle w_y, U^1 v^1_L \odot U^2 v^2_L \rangle + \langle V^1_y, v^1_L \rangle + \langle V^2_y, v^2_L \rangle + b_y - \log(Z),
\]

where \( \odot \) is the element-wise vector product. Hence, for fixed \((U^1, U^2)\), we are learning a linear hyperplane in the fused space of dimension \( F \). The projection on \( U^1 \) and \( U^2 \) defines a CCA-like lower dimensional spaces, where the two modalities are maximally correlated. The fused space is then defined as the element-wise vector product between two co-occurring vectors in the CCA-like lower dimensional spaces. Hence, we can think of the last layer of the bilinear, bimodal DNN as being an ordinary softmax, having the following input \((v^1_L, v^2_L, U^1 v^1_L \odot U^2 v^2_L) \in \mathbb{R}^{K^1 + K^2 + F}\). Therefore the decision function is learned based on the individual contributions of the modalities \( v^1_L \) and \( v^2_L \), as well as the joint representation produced by the fused space \( U^1 v^1_L \odot U^2 v^2_L \).

### 6.4.2 Factored Bilinear Softmax With Sharing

When the classes we would like to predict are organized as the leaves of a tree structure of depth two, we can further reduce the computational complexity by sharing weights between leaves having the same parent node. This is the case in AV-ASR as the 1328 contextual phoneme states are organized as leaves of a tree, where the parent nodes correspond to 42 different phoneme categories. In that case we share the bilinear term across leaves having the same parents. By doing so in the case of AV-ASR, we are only taking into account the correlations between the audio and the visual channel at the phoneme level, rather than on a fine grained grid of contextual states. We can think of this sharing as a pooling operation at the phoneme level. More formally, assume that the label set \( \mathcal{Y} \) is partitioned into \( G \) non
overlapping groups \( \{ \mathcal{Y}_g \}_{g=1\ldots G} \), we assume that:

\[
W^y = W^g = U^1 \text{diag}(w_g) U^{2,\top}, \forall y \in \mathcal{Y}_g, g = 1\ldots G.
\]

Hence we reduce the number of weights to learn for the joint representation from \( C \times F \) to \( G \times F \).

### 6.5 Back-propagation with the Factored Bilinear DNN with Sharing

In this section we give the back-propagation algorithm and the update rules for the bilinear DNN with sharing (\( b^2 DDN-wS \)). Recall that our classes have a tree structure with leaves \( y \), and parent nodes \( g \); a training example is therefore labeled by its leave label \( y \) (States) as well its parent node \( g \) (Phonemes), \( (x^1, x^2, y, g), y \in \{ 1\ldots C \}, \) and \( g \in \{ 1\ldots G \} \). We use the notation \( g(y) \) to note the group to which \( y \) belongs, and we set \( \text{Root}_{g(y)} = 1, \text{Root}_g = 0, g = 1\ldots G, g \neq g(y) \).

For the bilinear softmax with sharing, we keep track of the errors at the level of the labels (States), as well as the groups level (Phonemes):

\[
\delta^k_L = t_k - \rho(k | v^1_L, v^2_L), \quad k = 1\ldots C, \quad \delta_L \in \mathbb{R}^{C \times 1}.
\]

\[
\delta^g_G = \text{Root}_g - \sum_{k \in \mathcal{Y}_g} \rho(k | v^1_L, v^2_L), \quad g = 1\ldots G, \quad \delta_G \in \mathbb{R}^{G \times 1}.
\]

Let \( W = [w_1, \ldots, w_G] \in \mathbb{R}^{F \times G} \), the gradients of the parameters of the bilinear softmax are given by:

\[
\frac{\partial \mathcal{E}}{\partial W} = (U^{1,\top} v^1_L \odot U^{2,\top} v^2_L) \delta^\top_G.
\]

\[
\frac{\partial \mathcal{E}}{\partial U^1} = v^1_L v^{2,\top} U^2 \text{diag}(W \delta_G), \quad \frac{\partial \mathcal{E}}{\partial U^2} = v^2_L v^{1,\top} U^1 \text{diag}(W \delta_G).
\]

\[
\frac{\partial \mathcal{E}}{\partial V^1} = v^1_L \delta^\top_L, \quad \frac{\partial \mathcal{E}}{\partial V^2} = v^2_L \delta^\top_L, \quad \frac{\partial \mathcal{E}}{\partial b} = \delta_L.
\]

For the layer right before the Bilinear softmax, we have a double projection to the first modality network (audio stream) and to the second modality network (visual stream).

We need to compute:

\[
W^g = U^1 \text{diag}(w_g) U^{2,\top}, g = 1\ldots G.
\]

\[
m^{2\rightarrow 1}_g = W^g v^2_L \quad M^{2\rightarrow 1}_L = [m_1^{2\rightarrow 1}, \ldots, m^2_G \rightarrow 1] \in \mathbb{R}^{K^1_L \times G}.
\]

\[
m^{1\rightarrow 2}_g = W^g \cdot v^1_L \quad M^{1\rightarrow 2}_L = [m_1^{1\rightarrow 2}, \ldots, m^1_G \rightarrow 2] \in \mathbb{R}^{K^2_L \times G}.
\]
Let $V^j = [V_1^j \ldots V_C^j]$, $j \in \{1, 2\}$, hence the errors we propagate to each network have the following form:

\[
\delta_{L-1}^j = \frac{\partial E}{\partial v_{L-1}^j} = M_{L-1}^{2,1} \delta_G^1 + V^1 \delta_L^j. \tag{6.4}
\]

\[
\delta_{L-1}^2 = \frac{\partial E}{\partial v_{L-1}^2} = M_{L-1}^{1,2} \delta_G^2 + V^2 \delta_L^j. \tag{6.5}
\]

Note that the errors now have an additional term, $M_{L-1}^{2,1} \delta_G^1$, and $M_{L-1}^{1,2} \delta_G^2$, respectively. We can think of those terms as messages passed between networks through the bilinear term. In that way, one network influences the weights of the other one. For the rest of the updates, it follows standard back-propagation in both networks; we give it here for completeness.

Let $u_1^\ell = W_1^\ell v_1^\ell + b_1^\ell$, $u_2^\ell = W_2^\ell v_2^\ell + b_2^\ell$, then finally we have:

\[
\frac{\partial E}{\partial W_1^\ell} = v_1^\ell (\text{diag}(\sigma'(u_1^\ell) \delta_L^1)^\top, \frac{\partial E}{\partial b_1^\ell} = \text{diag}(\sigma'(u_1^\ell)) \delta_L^1,
\]

\[
\delta_{L-1}^j = W_j^\ell \delta_L^j, \quad j \in \{1, 2\}, \quad \ell = L - 1 \ldots 1,
\]

where $\delta_{L-1}^1$, and $\delta_{L-1}^2$ are given in equations (6.4) and (6.5). For each variable $\theta$, we have an update rule $\theta \leftarrow \theta + \eta \frac{\partial E}{\partial \theta}$, where $\eta$ is the learning rate. For $U^1$ and $U^2$, we need to keep control of the Frobenius norm by following the gradient step with a projection to the Frobenius ball $U^j \leftarrow U^j \min(1, \frac{\lambda}{||U^j||_F})$, $j \in \{1, 2\}$.

**Remark 1.** For the bilinear softmax without sharing the update rules are similar ($\delta_G$ is replaced by $\delta_L$).

### 6.6 Combining Posteriors from Bimodal and Bilinear Bimodal Networks

We experiment with various factored $b^2$-DNN-$wS$ architectures, initialized at random on the IBM AV-ASR Large Vocabulary Studio Dataset. We use the following notation for the architecture of the bilinear network: \([arch_a [arch_v] F]\), where $arch_a$ and $arch_v$ are the architectures of the audio and the visual network respectively, and $F$ is the dimension of the fused space.

We consider architectures by increasing complexity $Arch = [360, 500, 500, 200, 1328|540, 500, 500, 200, 1328|F = 200], \ Arch_1 = [360, 600, 600, 400, 100, 1328|540, 600, 600, 400, 100, 1328|F = 100], \ Arch_2 = [360, 500, 500, 500, 500, 500, 500, 200, 1328|540, 500, 500, 500, 500, 500, 200, 1328|F = 200]$. In all our experiments we set $\lambda = 2$. Recall that the bimodal DNN using the separate training paradigm introduced in Section 6.3 achieves 35.83% PER. As shown in Table 6.2, each architecture alone does not improve on the bimodal DNN, but averaging the posteriors of the three architectures we obtain a small gain. A gain of 1.8% absolute is obtained by averaging the posteriors of the bimodal and the bilinear bimodal networks,
showing that the bilinear networks have uncorrelated errors with the bimodal network.

<table>
<thead>
<tr>
<th></th>
<th>PER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arch</td>
<td>38.89%</td>
</tr>
<tr>
<td>Arch₁</td>
<td>39.01%</td>
</tr>
<tr>
<td>Arch₂</td>
<td>38.36%</td>
</tr>
<tr>
<td>Bimodal</td>
<td>35.83%</td>
</tr>
<tr>
<td>Arch + Arch₁ + Arch₂</td>
<td>35.54%</td>
</tr>
<tr>
<td>Arch + Arch₁ + Arch₂ + Bimodal</td>
<td>34.03%</td>
</tr>
</tbody>
</table>

Table 6.2: Empirical evaluation on the AV-ASR Studio dataset.

6.7 Conclusion

In this chapter we have studied deep multimodal learning for the task of phonetic classification from audio and visual modalities. We demonstrate that even in clean acoustic conditions using visual channel in addition to speech results in significantly improved classification performance. A bilinear bimodal DNN is introduced which leverages correlation between the audio and visual modalities, and leads to further error rate reduction.
Part II

Q-ary Compressive Sensing and Embedding
Chapter 7

Q-ary Compressive Sensing

We introduce in this Chapter $q$-ary compressive sensing, an extension of 1-bit compressive sensing. We propose a novel sensing mechanism and a corresponding recovery procedure. The recovery properties of the proposed approach are analyzed both theoretically and empirically. Results in 1-bit compressive sensing are recovered as a special case. Our theoretical results suggest a tradeoff between the quantization parameter $q$, and the number of measurements $m$ in the control of the error of the resulting recovery algorithm.

7.1 Introduction

An embedding is a map $f$ from a metric space $(\mathcal{X}, d_\mathcal{X})$ to another metric space $(\mathcal{Y}, d_\mathcal{Y})$, such that:

$$d_\mathcal{Y}(f(x), f(x')) \sim d_\mathcal{X}(x, x') \forall x, x' \in \mathcal{X}.$$ 

Embedding a set of high dimensional signals into a lower dimensional space, in a similarity preserving manner, has found many applications in fast and efficient signals acquisition and recovery [DDT+08], as well as in dimensionality reduction and fast information retrieval in many machine learning applications [Ind01].

The so-called Johnson-Lindenstrauss Transform [JL84] illustrates the power of random projections in embedding a set of $N$ points into a lower dimensional space of dimension $m = O(\log(N))$ while preserving approximately pairwise Euclidean distances. At the heart of compressive sensing [CT06], the restricted isometry property of random matrices, shows that when restricted to the set of $s$-sparse signals in $\mathbb{R}^n$, the Euclidean distance is approximately preserved between two signals from only $m = O(s \log(n))$ random projections. More generally for a set $K \subset \mathbb{R}^n$ characterized by a certain measure of complexity $w(K)$, namely the Gaussian mean width of the set,

$$w(K) = \mathbb{E} \sup_{x \in K, y \in K} |\langle g, x - y \rangle|, g \sim \mathcal{N}(0, I_n),$$
we have a similar restricted isometry property from only \( m = O(w^2(K)) \) random projections \([KM05]\). We can think therefore of random projections with a random matrix \( A \in \mathbb{R}^{m \times n} \) as being a coding \( C \):

\[
C: (K, ||.||_2) \mapsto (\mathbb{R}^m, ||.||_2)
\]

\[
x \mapsto C(x) = Ax.
\]

such that \( ||C(x) - C(x')||_2 \sim (1 \pm \epsilon)||x - x'||_2 \), for all \( x, x' \in K \), where \( 0 < \epsilon < 1 \) and \( m \) depends on a measure of the complexity of the set \( K \).

Since \( m \) is typically smaller than \( n \), the mapping \( C \) is versatile, it can be seen as a quasi-isometry and a dimensionality reduction that allows fast distances computation, but also as a lossy compression of the signals in \( K \). Hence the natural question raised in the field of compressive sensing \([CT06]\) consists in defining a decoder \( D \):

\[
D: (\mathbb{R}^m, ||.||_2) \mapsto (K, ||.||_2)
\]

\[
b = C(x_0) \mapsto D(b).
\]

An exact recovery is obtained if, for a given \( x_0 \in K \subset \mathbb{R}^n, D(C(x_0)) = x_0 \), and \( \epsilon - \) recovery is obtained if \( ||D(C(x_0)) - x_0||_2 \leq \epsilon \).

Let \( b = Ax_0 \), in the noiseless setting we can define \( D \) as:

\[
\text{find } x, \quad \text{subject to } \quad Ax = b \quad x \in K. \tag{7.1}
\]

In the noisy setting, \( D \) can be defined as follows:

\[
\arg \min_{x \in K} ||Ax - b||^2 \tag{7.2}
\]

For a convex set \( K \), problems (7.1),(7.2), can be solved using convex programing. For non convex sets, a convex relaxation approach \([CRPW12]\) consists in replacing \( K \) by the convex hull of \( K \).

To sum up consider a random Gaussian matrix \( A \in \mathbb{R}^{m \times n} \), \( C(x) = Ax \) defines an embedding from \( (K \subset \mathbb{R}^n, ||.||_2) \) to \( (\mathbb{R}^m, ||.||_2) \), where \( m \) depends on a measure of the complexity of the set \( K \). As mentioned earlier \( C \) defines a new paradigm in the acquisition of structured signals known as compressive sensing, and in fast computation of approximate pairwise distances known as locality sensitive hashing \([IN07, AI08]\).

Quantizing random projections is a step forward in bypassing the classic pipeline of analogue acquisition of linear measurements followed by scalar quantization techniques and substituting it by a compressive acquisition of non linear quantized measurements, such as
the so called one bit measurements $C(x) = \text{sign}(Ax)$ [BB08]. For instance $C$ defines an embedding from a set $K$ on the unit sphere $S^{n-1}$, $(K \subset S^{n-1}, d_G)$ to the binary Hamming cube $(\{-1, 1\}^m, d_H)$, where $d_G$ is the geodesic distance on the sphere and $d_H$ is the Hamming distance (i.e. for $a, b \in \{-1, 1\}^m$, $d_H(a, b) = \frac{1}{m} \sum_{i=1}^{m} 1_{a_i \neq b_i}$),

$$C: (K \subset S^{n-1}, d_G) \mapsto (\{-1, 1\}^m, d_H)$$
$$x \mapsto C(x) = \text{sign}(Ax).$$

so that : $d_H(C(x), C(x')) \sim d_G(x, x') \pm \delta$, for all $x, x' \in K$, where $0 < \delta < 1$, and $m$ depends on a measure of complexity of the set $K$ [PV11, JLBB13].

Since $m$ is typically small and the Hamming distance can be computed efficiently, this embedding is widely used in approximate nearest neighbor search and locality sensitive hashing. On the other hand given $y = \text{sign}(Ax_0)$, a natural question raised in the growing field of one bit compressive sensing [BB08, PV13b, PV13a, JLBB13], consists in recovering $x_0$, i.e. in finding a decoder $D$, such that $\|D(C(x_0)) - x_0\| < \epsilon$:

$$D: (\{-1, 1\}^m, d_H) \mapsto (K \subset S^{n-1}, \|\|_2)$$
$$y = C(x_0) \mapsto D(y).$$

We can define $D$ [BB08, PV13a] as:

$$\text{find } x, \quad \text{subject to } y_i(Ax)_i \geq 0 \quad i = 1 \ldots m \quad x \in K \cap S^{n-1}. \quad (7.3)$$

The problem in (7.3) is non convex, if $K$ was a convex set, in [PV13b] it is shown that the following convex relaxation allows to find efficiently an $\epsilon$- recovery from $O(w^2(K))$ one bit measurements:

$$\text{arg max } \frac{1}{m} \sum_{i=1}^{m} y_i (a_i, x) \quad (7.4)$$

where $\mathbb{B}^n$ is the unit ball in $\mathbb{R}^n$. We note that this formulation is similar to binary classification in term of the margin (or the correlation) maximization. If $K$ was not convex we replace it by its convex hull. We note that one bit measurements are robust to increasing (on average ) non linear and unknown distortion or noise [PV13b]. If instead of observing $y = \text{sign}(\langle a, x_0 \rangle)$, we observe $y = \theta (\langle a, x_0 \rangle)$, where $\theta$ is a non linearity such that:

$$\lambda = \mathbb{E}(g \theta(g)) > 0, \quad g \sim \mathcal{N}(0, 1), \quad (7.5)$$

it is shown in [PV13b] that the recovery is still possible. Let $\hat{x}_m$ be the solution of (7.4) [PV13b] states that:

For $m \geq \frac{Cw^2(K)}{\epsilon^2}$, $\|\hat{x}_m - x_0\|^2 \leq \frac{\epsilon}{\lambda}$, with probability $1 - 8 \exp(-c\epsilon^2 m)$. 

82
To stress the role played by $\lambda$, authors in [PV13b] consider the following two cases:

- The noiseless model: $\theta(z) = \text{sign}(z)$, $\lambda$ achieves its maximum value $\lambda = \sqrt{\frac{2}{\pi}}$.
- The noisy model: $\theta(z) = \text{sign}(z + \nu)$ where $\nu \sim N(0, \sigma^2)$, $\lambda = \sqrt{\frac{2}{\pi(1+\sigma^2)}}$.

We see that $\lambda$ is decreasing with the level of noise hence it plays the role of a signal to noise ratio (SNR).

In this Chapter we ask the question of extending the ideas of binary embeddings to a $q$-ary alphabet embedding. Getting compact codes defined on a $q$-ary alphabet representing high dimensional data in a similarity preserving manner allows us to get a finer representation of the data, and to design new locality sensitive hashing schemes for approximate near neighbor search. In addition we would like to be able also to define a pair of coder/decoder $(C,D)$:

$$C : K \cap S^{n-1} \rightarrow \{0 \ldots q - 1\}^m,$$

$$D : \{0 \ldots q - 1\}^m \rightarrow K \cap S^{n-1},$$

to keep close to the spirit of the binary case previously discussed.

The coding $C$ defines a $q$-ary embedding and the decoder $D$ defines a compressive sensing framework that we call $q$-ary compressive sensing.

### 7.2 Q-ary Compressive Sensing

In this section we first describe the sensing and recovery procedure in Section 7.2.1, then describe the recovery results in Section 7.3. We give an experimental validation in Section 7.4 and finally we give our theoretical analysis in Section 7.6.

#### 7.2.1 Sensing and Recovery

The sensing procedure we consider is given by a map $C$ from $\mathbb{R}^n$ to the $q$-ary, $m$ Hamming cube $\{0, \ldots, q - 1\}^m$. To define $C$ we recall the following definition of the simplex coding used in the multiclass setting introduced in Chapter 2.

**Definition 9** (Simplex Coding). *The simplex coding map is $S : \{0, \ldots, q - 1\} \rightarrow \mathbb{R}^{q-1}$, $S(j) = c_j$, where*

1) $\|c_j\|^2 = 1$,

2) $\langle c_j, c_i \rangle = -\frac{1}{q-1}$, for $i \neq j$,

3) $\sum_{j=0}^{q-1} c_j = 0$. 
Definition 10 (q-ary Quantized Measurements). Let $W \in \mathbb{R}^{q-1,n}$ be a Gaussian random matrix, i.e. for all $i, j, W_{ij}$ are iid, $W_{ij} \sim \mathcal{N}(0, 1)$. For $x_0 \in \mathbb{R}^n$,

$$Q : \mathbb{R}^n \to \{0, \ldots, q - 1\}, \quad Q(x_0) = Q_W(x_0) = \arg \max_{j=0 \ldots q-1} (c_j, Wx_0),$$

is called a q-ary quantized measurement.

Then, we can define the q-ary sensing strategy induced by non linear quantized measurements.

Definition 11 (q-ary Sensing). Let $x_0 \in \mathbb{R}^n$. Let $W_1, \ldots, W_m$, be independent Gaussian random matrices in $\mathbb{R}^{q-1,n}$ and $Q_{W_i}(x_0), i = 1, \ldots, m$ as in Def. 10. The q-ary sensing is $C : \mathbb{R}^n \to \{0, \ldots, q - 1\}^m$,

$$C(x_0) = (Q_{W_1}(x_0), \ldots Q_{W_m}(x_0)).$$

Remark 2. Note that if $q = 2$, $W$ reduces to a Gaussian random vector, and $2Q(x_0) - 1 = \text{sign}(Wx_0)$, so that the q-ary quantized measurements become equivalent to those considered in 1-bit CS.

In this Chapter, we are interested in recovering a signal $x_0$ from its q-ary measurements $y = (y_1, \ldots, y_m) = C(x_0)$.

Since $\sum_{j=0 \ldots q-1} c_j = 0$, it is easy to see that for all $w \in \mathbb{R}^{q-1}$, we have $\max_{j=0 \ldots q-1} (c_j, w) \geq 0$.

Hence we can write the following recovery problem:

$$\text{find } x, \quad \text{subject to } \langle c_{y_i}, W_i x \rangle \geq 0, \quad i = 1 \ldots m, \quad \|x\|_2^2 = 1. \quad (7.6)$$

Indeed, as in one-bit compressive sensing, we cannot hope to recover the norm of the vector from inequality constraints, hence the norm one constraint.

Following [PV13b], we consider the following relaxation:

$$\max_x \quad \frac{1}{m} \sum_{i=1}^m \langle c_{y_i}, W_i x \rangle, \quad \text{subject to } \|x\|_2^2 = 1. \quad (7.7)$$

If we further more assume that $x_0$ belongs to a set $K \subset \mathbb{R}^n$, we can consider the following recovery strategy $D : \{0, \ldots, q - 1\}^m \to K \cap S^{n-1}$ defined by,

$$D(y) = \arg \max_{x \in K \cap S^{n-1}} \frac{1}{m} \sum_{i=1}^m \langle c_{y_i}, W_i x \rangle. \quad (7.8)$$

When $K$ is a convex set in $\mathbb{R}^n$, we follow [PV13b] in replacing the non convex norm one constraint $x \in S^{n-1}$ by a convex inequality constraint $\|x\|_2^2 \leq 1$, i.e $x \in \mathbb{B}^n$, and obtain the
following convex relaxation:

\[ D(y) = \arg \max_{x \in K \cap \mathbb{R}^n} \frac{1}{m} \sum_{i=1}^{m} \langle c_{y_i}, W_i x \rangle, \quad (7.9) \]

let

\[ \hat{x}_m = D(y) = D(C(x_0)), \]

be the solution of Problem (7.9).

**Remark 3 (Connection to Classification).** An inspiration for considering the q-ary CS stems from an analogy between 1-bit compressed sensing and binary classification in machine learning. In this view, Definition (11) is related to the approach proposed for multicategory classification in this thesis. Following these ideas, we can extend the recovery strategy (7.9) by considering

\[ D_{V}(y) = \arg \min_{x \in K \cap \mathbb{R}^n} \frac{1}{m} \sum_{i=1}^{m} V(\langle c_{y_i}, W_i x \rangle), \quad (7.10) \]

where \( V \) is a convex, L Lipschitz, decreasing loss function \( V : \mathbb{R} \rightarrow \mathbb{R}^+ \). Problem (7.9) corresponds to the choice \( V(x) = -x \). Other possible choices include \( V(x) = \max(1 - x, 0) \), \( V(x) = \log(1 + e^{-x}) \), and \( V(x) = e^{-x} \). All these loss functions can be seen as convex relaxations of the 0-1 loss function, defined as \( V(x) = 0 \) if \( x \leq 0 \), and \( 1 \) otherwise. The latter defines the misclassification risk, which corresponds to Hamming distance in Compressive Sensing, which is the natural measure of performance while learning classification rules.

### 7.2.2 The signal Set & the Gaussian Mean Width

In this section we describe the notion of Gaussian mean width to characterize a set \( K \subset \mathbb{R}^n \). For a succinct review we refer the reader to [PV13b], and references therein. The Gaussian mean width of a set \( K \subset \mathbb{R}^n \) is defined as follows:

\[ w(K) = \mathbb{E} \sup_{x \in K, y \in K} |\langle g, x - y \rangle|, \quad g \sim \mathcal{N}(0, I_n). \]

Intuitively the gaussian mean width \( w(K) \) is proportional to the diameter of the set \( K \), that we note by \( \text{diam}(K) \), and in fact we have:

\[ \sqrt{\frac{2}{\pi}} \text{diam}(K) \leq w(K) \leq n^{\frac{1}{2}} \text{diam}(K). \quad (7.11) \]

Note that for an origin-symmetric set \( K \), we have:

\[ w(K) = 2\mathbb{E} \left( \sup_{x \in K} |\langle g, x \rangle| \right). \]
7.3 Recovery guarantees

The following theorem describes the recovery guarantees for the proposed procedure, when applied on a signal $x_0$, belonging to an origin symmetric set $K \subset \mathbb{R}^n$. We note by

$$w(K \cap \mathbb{B}^n),$$

the gaussian mean width of the intersection of the set $K$, with the $\ell_2$ unit ball $\mathbb{B}^n$.

**Theorem 7** (Q-ary CS Recovery Guarantees). Let $x_0 \in K \subset \mathbb{R}^n, \|x\|_0 = 1$, let $y = C(x_0)$, the q-ary measurements as defined in Definition 11. Let $\delta > 0$, and $m \geq C\delta^{-2}w^2(K \cap \mathbb{B}^n)$. Then with probability at least $1 - 4\exp(-c\delta^2 m)$ (where $C > 0$ and $c > 0$ are universal constants), the solution $\hat{x}_m = D(y)$ of problem (7.9) satisfies,

$$\|\hat{x}_m - x_0\|_2^2 \leq \frac{\delta}{\sqrt{\log(q)}}.$$  \hspace{1cm} (7.12)

**Proof of Theorem 7.** The proof of Theorem 7 is given in Section 7.6. \hfill \Box

Here, we add four comments. First, we note that the above result implies the error bound,

$$\|\hat{x}_m - x\|_2^2 \leq C \left( \frac{w(K \cap \mathbb{B}^n)}{\sqrt{\log(q)m}} + \beta \right),$$  \hspace{1cm} (7.13)

with probability at least, $1 - 4\exp(-c\beta^2 m), \beta > 0$, for certain universal constants $c, C > 0$.

Second, Inequalities (7.12), (7.13) can be compared to results in 1-bit CS. For the same number of measurements, $m \geq C\delta^{-2}w^2(K \cap \mathbb{B}^n)$, the error for q-ary CS is $\frac{\delta}{\sqrt{\log(q)m}}$, in contrast with $\frac{\delta}{\sqrt{\log(q)}}$ in the 1-bit CS [PV13b], at the expense of a more demanding sensing procedure. Also note that, for $q = 2$, we recover the result in 1-bit CS as a special case.

Third, we see that for a given accuracy our results highlights a trade-off between the number of q-ary measurements $m$ and the quantization parameter $q$. To achieve an error $\epsilon$, with a memory budget of $\ell$ bits, one can choose $m$ and $q$ so that $\epsilon = O\left(\frac{1}{\sqrt{m \log(q)}}\right)$, and $m \log_2(q) = \ell$ (see also section 7.4.2).

Finally, in the following we will be interested in $K$ being the set of approximately $s$-sparse signals (the $\ell_1$ ball relaxation). Following again [PV13b], it is interesting to consider in Problem (7.9) the relaxation

$$K_1 = \{u \in \mathbb{R}^n : \|u\|_1 \leq \sqrt{s}\}.$$
With this choices, it is possible to prove that \( w(K_1 \cap B^n) \leq C \sqrt{s \log \left( \frac{2n}{s} \right)} \), and that for \( m \geq C \delta^{-2} s \log \left( \frac{2n}{s} \right) \), the solution of the convex program (7.9) on \( K_1 \) satisfies, \( \| \hat{x}_m - x \|^2 \leq \frac{\delta}{\sqrt{\log(q)}} \).

We end noting that other choices of \( K \) are possible, for example in [RNR11] the set of group sparse signals (and their Gaussian width) are studied.

### 7.4 Experimental Validation

In this section, we describe some numerical simulations in sparse recovery, Section 7.4.1, and preliminary experiments in an image recovery problem, Section 7.4.3.

#### 7.4.1 An Algorithm for Sparse recovery

In our experiments, we considered the following variation of problem (7.9), Let \( \xi_i = c^\top y_i W_i \in \mathbb{R}^n, i = 1 \ldots m. \)

\[
\max_{u, \|u\|_2 \leq 1} \frac{1}{m} \sum_{i=1}^{m} \langle \xi_i, u \rangle - \eta \|u\|_1, \tag{7.14}
\]

where \( \eta > 0. \) The above problem can be solved efficiently using Proximal Methods, a solution can be computed via the iteration,

\[
\begin{align*}
    u_{t+1} &= u_t + \nu_t \frac{1}{m} \sum_{i=1}^{m} \xi_i, \\
    u_{t+1} &= \text{Prox}_\eta(u_{t+1}), \\
    u_{t+1} &= u_{t+1} \min(1 - \frac{\eta}{\|u_{t+1}\|_2}, 1).
\end{align*}
\]

Where \( \nu_t \) is the gradient step size, and \( \text{Prox}_\eta \) acts component wise as \( \max(1 - \frac{\eta}{|u_i|}, 0)u_i. \) The iteration is initialized randomly to a unit vector.

**Remark 4.** The computational complexity of the sensing process depends on both \( m \) and \( q. \) Whereas, the computational complexity of the recovery algorithm, once computed \( \xi_i, \) is independent to the choice of \( q, \) and depends only on \( m \) and is the same as in 1-bit CS.

#### 7.4.2 Sparse Recovery

We tested our approach for recovering a signal from its \( q \)-ary measurements. We considered sparse signals of dimension \( d \) generated via a Gauss-Bernoulli model. In Figure 10-4(a), we see that the reconstruction error of \( \hat{x}_m \) (in blue), for varying \( q \) and \( m \) fixed, follows the theoretical bound \( \frac{1}{\sqrt{\log(q)}} \) (in red). In Figure 10-4(b), we see that the reconstruction error \( \hat{x}_m \) (in blue), for varying \( m \) and \( q \) fixed, follows the theoretical bound \( \frac{1}{\sqrt{m}} \) (in red). Figures 10-4(c), and 10-4(d) highlight the tradeoff between the number of measurements and the
quantization parameter. For a precision $\epsilon$, and a memory budget $2^B$, one can choose an operating point $(m, q)$, according to the theoretical bound \( \frac{1}{\sqrt{m \log(q)}} \).

![Figure 7-1: q-ary Compressive Sensing: Quantization/Number of measurements tradeoff.](image)

(a) Error $\|x_0 - \hat{x}_m\|^2$ versus $q$, for $m = 70, n = 100$.  
(b) Error $\|x_0 - \hat{x}_m\|^2$ versus $m$, for $q = 3, n = 100$.

(c) Theoretical bound for $\|x_0 - \hat{x}_m\|^2$ versus $m$ and $q$.

(d) Error $\|x_0 - \hat{x}_m\|^2$ versus $m$ and $q$.

Figure 7-1: q-ary Compressive Sensing: Quantization/Number of measurements tradeoff.

### 7.4.3 Image Reconstruction

Then, we considered the problem of recovering an image from $q$-ary measurements. We used the standard 8-bit grayscale boat image of size $64 \times 64$ pixels shown in Figure 7-3(a). We extracted the wavelet coefficients and performed thresholding to get a sparse signal. We normalized the resulting vector of wavelet coefficients of dimension $n = 3840$ to obtain a unit vector. Then, we performed sensing and recovery with $q = 2^5$ (5-bit compressive sensing) and $q = 2$ (1-bit compressive sensing) for the same $m = 2048 < n$ (See Figure 7-2 for details on the sensing mechanism). We compared the SNR performances of the
corresponding reconstructed images in a noiseless setting (Figures 7-3(b)-(c)), and a noisy setting, considering the noise before quantization model: 

\[ y_i = \arg \max_{j=0...q-1} \langle c_j, W_i x_0 + \nu_i \rangle, \]

where \( \nu_i \sim \mathcal{N}(0, \sigma I_{q-1}) , \text{iid} \ i = 1 \ldots m \) with \( \sigma = 0.8 \) (Figures 7-3(d)-(e)).

![Diagram](image)

Figure 7-2: \( Q \)-ary sensing for \( q = 3 \).

The results confirm our theoretical finding: higher quantization improves the SNR of the recovery.

## 7.5 Conclusion

In this Chapter we introduced \( q \)-ary compressive sensing an extension of one bit compressive sensing, and showed that there is a tradeoff between the quantization level, and the number of measurements in governing the recovery error. In the next chapter we answer the following
Figure 7-3: Image recovery with q-ary CS. (a) Original image. (b) Reconstruction with no-noise: \( q = 2^5 \), SNR = 20.2 dB. (c) Reconstruction with no-noise: \( q = 2 \), SNR = 16.2 dB. (d) Reconstruction with noise: \( q = 2^5 \), SNR = 18.3 dB. (e) Reconstruction with noise: \( q = 2 \), SNR = 15 dB.

question:

What are the properties of the q-ary embedding before and after quantization? What kind of similarity does the un-quantized q-ary sensing provide us? and what is the benefit of the quantized q-ary sensing in locality sensitive hashing and nearest neighbor search?

7.6 Theoretical Analysis

In this Section we give the proof of Theorem 7, we start first by a preliminary definition:

**Definition 12** (Risk and Empirical Risk). Let \( x_0 \in \mathbb{R}^n \), such that \( ||x||_0 = 1 \), and let \( W \sim \mathcal{N}(0, I_{(q-1)\times n}) \). For \( x \in \mathbb{R}^n \), \( ||x|| \leq 1 \), define

\[
\mathcal{E}^{x_0}(x) = \mathbb{E}_W (\langle y, Wx \rangle ),
\]

where \( y = Q_W(x_0) \).
Let $W_1 \ldots W_m$ be $m$ random independent gaussian matrices in $\mathbb{R}^{q-1,d}$, define
\[
\tilde{\mathcal{E}}^{x_0}(x) = \frac{1}{m} \sum_{i=1}^{m} \langle c_{y_i}, W_i x \rangle,
\]
where $y = (y_1, \ldots, y_m) = C(x_0)$. Let $\tilde{x}_m = D(y)$, as defined in equation (7.9).

In order to prove Theorem 7, we need a comparison inequality in expectation given in Lemma 7, and then using tools from concentration inequalities, and empirical processes theory we obtain the proof of Theorem 7:

**Lemma 7 (Comparison Inequality).** For any $x \in \mathbb{B}^n, x_0 \in \mathbb{S}^{n-1}$, the following inequality holds:
\[
\frac{1}{2} \|x - x_0\|^2 \leq \frac{1}{\lambda(q)} \left( \mathcal{E}^{x_0}(x_0) - \mathcal{E}^{x_0}(x) \right),
\]
where $\lambda(q) = \mathbb{E}(\xi), \xi = \max_{j=1 \ldots q} \langle c_j, g \rangle, g \sim \mathcal{N}(0, I_{q-1})$. There exists an absolute constant $c$ such that:
\[
\frac{1}{2} \|x - x_0\|^2 \leq \frac{c}{\sqrt{\log(q)}} \left( \mathcal{E}^{x_0}(x_0) - \mathcal{E}^{x_0}(x) \right),
\]

**Proof of Lemma 7.** Note that $W x_0 \sim \mathcal{N}(0, I_{q-1})$, and $W x \sim \mathcal{N}(0, I_{q-1})$. Hence $[W x_0, W x]$ is a jointly gaussian vector with cross covariance terms equal to $\langle x_0, x \rangle$. Therefore there exists $g \sim \mathcal{N}(0, I_{q-1})$ and $h \sim \mathcal{N}(0, I_{q-1})$, such that $g$ and $h$ are independent and:
\[
W x_0 = g \quad \text{and} \quad W x = \langle x_0, x \rangle g + (\|x\|^2 - \langle x_0, x \rangle^2)^{\frac{1}{2}} h.
\]

\[
\mathcal{E}^{x_0}(x_0) - \mathcal{E}^{x_0}(x) = \mathbb{E} \left( \langle c_y, W x_0 \rangle - \langle c_y, W x \rangle \right)
= \mathbb{E} \left( \langle c_y, g \rangle (1 - \langle x_0, x \rangle) + \langle c_y, h(\|x\|^2 - \langle x_0, x \rangle^2)^{\frac{1}{2}} \rangle \right)
= (1 - \langle x, x_0 \rangle) \mathbb{E} \left( \langle c_y, g \rangle \right),
\]
where the last equality follows from the fact that $y = \arg \max_{j=0 \ldots q-1} \langle c_j, g \rangle$ and $h$ are independent, and $\mathbb{E}(h) = 0$.

For $x \in \mathbb{B}^n, x_0 \in \mathbb{S}^{n-1}$, we have:
\[
1 - \langle x, x_0 \rangle \geq \frac{\|x - x_0\|^2}{2}.
\]

It follows that:
\[
\frac{1}{2} \|x - x_0\|^2 \leq \frac{1}{\lambda(q)} \left( \mathcal{E}^{x_0}(x_0) - \mathcal{E}^{x_0}(x) \right),
\]
where $\lambda(q) = \mathbb{E}(\langle c_y, g \rangle) = \mathbb{E}(\max_{j=0 \ldots q-1} \langle c_j, g \rangle)$.

**Bounding $\lambda(q)$.** Let $\eta_j = \langle c_j, g \rangle, \eta_j$ are dependent centered standard gaussian variables
with cross covariance term \( E(\eta_j \eta_k) = -\frac{1}{q-1} \). Let \( \nu_j, j = 1 \ldots q \), be \( q \), independent standard centered gaussian random variables.

\[
E(\eta_j - \eta_k)^2 = 2 - 2E(\eta_j \eta_k) = E(\nu_j - \nu_k)^2 + \frac{2}{q-1}. \tag{7.19}
\]

It follows from equation (7.19) that: \( E(\eta_j - \eta_k)^2 \geq E(\nu_j - \nu_k)^2 \). Then by the Slepian’s Inequality we have that:

\[
\lambda(q) = E(\max_j \eta_j) \geq E(\max_j \nu_j). \tag{7.20}
\]

By Sudakov minoration, for iid gaussians we have:

\[
E(\max_j \nu_j) \geq C \sqrt{\log(q)}, \tag{7.21}
\]

where \( C \) is an absolute constant. The bound with \( \sqrt{\log(q)} \) is known to be a tight bound, in fact \( E(\max_j=1...q \nu_j) \sim \sqrt{\log(q)} \). By combining (7.20) and (7.21) we conclude that:

\[
\frac{1}{\lambda(q)} \leq \frac{1}{C \sqrt{\log(q)}}. \tag{7.22}
\]

We have finally :

\[
\frac{1}{2} \| x - x_0 \|^2 \leq \frac{1}{C \sqrt{\log(q)}} \left( \mathcal{E}^{x_0}(x_0) - \mathcal{E}^{x_0}(x) \right). \tag{7.23}
\]

We are now ready to give the proof of Theorem 7.

**Proof of Theorem 7.** Let \( K \) be an origin symmetric set in \( \mathbb{R}^n \). Let \( x_0 \in S^{d-1} \cap K, y = C(x_0) \), and \( \hat{x}_m = D(y) \), as defined in equation (7.9). We have by Lemma 7:

\[
\frac{1}{2} \| \hat{x}_m - x_0 \|^2 \leq \frac{1}{\lambda(q)} \left( \mathcal{E}^{x_0}(x_0) - \mathcal{E}^{x_0}(\hat{x}_m) \right). \tag{7.24}
\]

The right hand side of equation (7.24) can be bounded as follows:

\[
Z = \mathcal{E}^{x_0}(x_0) - \mathcal{E}^{x_0}(\hat{x}_m) = \left( \mathcal{E}^{x_0}(x_0) - \hat{\mathcal{E}}^{x_0}(x_0) \right) + \left( \hat{\mathcal{E}}^{x_0}(x_0) - \hat{\mathcal{E}}^{x_0}(\hat{x}_m) \right) + \left( \hat{\mathcal{E}}^{x_0}(\hat{x}_m) - \mathcal{E}^{x_0}(\hat{x}_m) \right) \leq 2 \sup_{x \in K \cap \mathbb{B}^n} \left| \hat{\mathcal{E}}^{x_0}(x) - \mathcal{E}^{x_0}(x) \right|, \tag{7.25}
\]

where the last inequality follows form the definition of \( \hat{x}_m \), as \( \hat{\mathcal{E}}^{x_0}(x_0) \leq \hat{\mathcal{E}}^{x_0}(\hat{x}_m) \).

We start by bounding the expectation of \( Z \), by finding a bound on the following expectation:

\[
\mu = \mathbb{E} \left( \sup_{x \in K \cap \mathbb{B}^n} \left| \hat{\mathcal{E}}^{x_0}(x) - \mathcal{E}^{x_0}(x) \right| \right). \]
Recall that:
\[
\hat{E}^{x_0}(x) = \frac{1}{m} \sum_{i=1}^{m} \langle \xi_i, x \rangle,
\]

where \( \xi_i = e_i^T W_i, \ i = 1 \ldots m. \)

Note that \( \xi_i \sim \mathcal{N}(0, I_n) \), are \( m \) iid Gaussian vectors. It follows by symmetrization, by introducing Rademacher variables \( \sigma_i \in \{ \pm 1 \}, i = 1 \ldots m \), \( \sigma_i \) are iid, such that \( \mathbb{P}(\sigma_i = 1) = \frac{1}{2} : \)

\[
\mu = \mathbb{E} \left( \sup_{x \in K \cap \mathbb{B}^n} \left| \hat{E}^{x_0}(x) - E^{x_0}(x) \right| \right) \leq 2\mathbb{E} \left( \sup_{x \in K \cap \mathbb{B}^n} \left| \frac{1}{m} \sum_{i=1}^{m} \sigma_i \langle \xi_i, x \rangle \right| \right)
\]

(7.26)

Using the rotation invariance and the symmetry of the Gaussian distribution we have:

\[
\frac{1}{m} \sum_{i=1}^{m} \sigma_i \xi_i \overset{\text{dist}}{=} \frac{1}{m} \sum_{i=1}^{m} \xi_i = \frac{1}{\sqrt{m}} \mathcal{N}(0, I_n) = \frac{1}{\sqrt{m}} w\ (\text{where } \xi \sim \mathcal{N}(0, I_n)).
\]

Therefore:

\[
\sup_{x \in K \cap \mathbb{B}^n} \left| \frac{1}{m} \sum_{i=1}^{m} \sigma_i \langle \xi_i, u \rangle \right| \overset{\text{dist}}{=} \frac{1}{\sqrt{m}} \sup_{x \in K \cap \mathbb{B}^n} |\langle \xi, x \rangle|.
\]

(7.27)

Since the set \( K \) is origin symmetric we have therefore:

\[
w(K \cap \mathbb{B}^n) = 2\mathbb{E} \left( \sup_{x \in K \cap \mathbb{B}^n} |\langle \xi, x \rangle| \right).
\]

Putting together (7.26) and (7.27) we have:

\[
\mu \leq \frac{w(K \cap \mathbb{B}^n)}{\sqrt{m}}.
\]

(7.28)

Taking expectation in (7.24),(7.25) and using (7.28) we get a first a result in expectation \( (\hat{x}(\xi_1, \ldots, \xi_m)):\)

\[
\frac{1}{2} \mathbb{E}(||\hat{x}_m - x_0||^2) \leq \frac{2w(K \cap \mathbb{B}^n)}{\lambda(q) \sqrt{m}} \leq 2c \frac{w(K \cap \mathbb{B}^n)}{\sqrt{\log(q)m}}.
\]

(7.29)

We turn now to see how well \( \sup_{x \in K \cap \mathbb{B}^n} |\hat{E}^{x_0}(x) - E^{x_0}(x)| \), concentrates around its expected value. The following deviation inequality holds, \( \forall \ t > 0 \) [PV13b]:

\[
\mathbb{P} \left\{ \sup_{x \in K \cap \mathbb{B}^n} \left| \hat{E}^{x_0}(x) - E^{x_0}(x) \right| \geq 2\mu + t \right\} \leq 4\mathbb{P} \left\{ \sup_{x \in K \cap \mathbb{B}^n} \left| \frac{1}{m} \sum_{i=1}^{m} \sigma_i \langle \xi_i, x \rangle \right| > \frac{t}{2} \right\}.
\]

Now Plugging (7.28) in the left hand side of the inequality and (7.27) in the right hand side
of the inequality, we obtain:

$$
P \left( \sup_{x \in K \cap B} \left| \mathcal{E}^{x_0}(x) - \mathcal{E}^{x_0}(x) \right| \geq \frac{2w(K \cap B)}{\sqrt{m}} + t \right) \leq 4P \left( \frac{1}{\sqrt{m}} \sup_{x \in K \cap B} |(\xi, x)| > \frac{t}{2} \right) . \quad (7.30)$$

Now using the gaussian concentration inequality [PV13b] we bound the right hand side of the above inequality:

$$
P\left( \sup_{x \in K \cap B} |(\xi, x)| \geq w(K \cap B) + r \right) \leq \exp \left( \frac{-r^2}{2} \right) . \quad (7.31)$$

Let $t = \frac{2w(K \cap B)}{\sqrt{m}} + \frac{2r}{\sqrt{m}}$, it follows from equations (7.30) and (7.31) that:

$$
P \left( \sup_{x \in K \cap B} \left| \mathcal{E}^{x_0}(x) - \mathcal{E}^{x_0}(x) \right| \geq \frac{4w(K \cap B)}{\sqrt{m}} + \frac{2r}{\sqrt{m}} \right) \leq 4 \exp \left( \frac{-r^2}{2} \right) .$$

Setting $\beta = \frac{2r}{\sqrt{m}}$, we obtain finally:

$$
P \left( \sup_{x \in K \cap B} \left| \mathcal{E}^{x_0}(x) - \mathcal{E}^{x_0}(x) \right| \geq \frac{4w(K \cap B)}{\sqrt{m}} + \beta \right) \leq 4 \exp \left( -\frac{\beta^2 m}{8} \right) \quad (7.32)$$

Hence, by equation (7.25) the following holds with a probability at least $1 - 4 \exp(-\frac{\beta^2 m}{8})$:

$$\mathcal{E}^{x_0}(x_0) - \mathcal{E}^{x_0}(\hat{x}_m) \leq \frac{8w(K \cap B)}{\sqrt{m}} + 2\beta,$$

Finally appealing to the comparison inequality (7.16), we have for any $\beta > 0$ with probability at least $1 - 4 \exp(-\frac{\beta^2 m}{8})$:

$$\frac{1}{2} \|\hat{x}_m - x_0\|^2 \leq \frac{c}{\sqrt{\log(q)}} \left( \frac{8w(K \cap B)}{\sqrt{m}} + 2\beta \right) .$$
Chapter 8

Random Max-Out Embedding

In this chapter, based on the q-ary sensing introduced in Chapter 7, we propose and study random max-out embedding or feature map for large scale learning. We show that the resulting random feature map, allows a locally linear estimation of the function of interest in a regression or a classification setting. The study of this random feature map gives insights on the recently introduced deep maxout network [GWFM+13], and more generally on Max-pooling used in diverse deep representations [ALR+13].

8.1 Introduction

We consider a classical supervised learning setting. Let $\mathcal{X} \subset \mathbb{R}^n$, be the input space and $\mathcal{Y} \subset \mathbb{R}$ be the output space (we consider for simplicity of the exposure single output space). We assume that we are given a training set $S = \{(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}, i = 1 \ldots N\}$.

Our purpose is to find a non linear function $f$ belonging to a hypothesis class of functions $\mathcal{H}$, that minimizes the empirical risk:

$$
\min_{f \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^{N} V(y_i, f(x_i)),
$$

where $V$ is a loss function. $V$ is a cross-entropy loss for instance.

In Kernel methods, functions are assumed to belong to a reproducing kernel Hilbert space (RKHS) $\mathcal{H}_K$, empirical risk minimization in that setup leads to a rich class of non linear algorithms, via regularization in RKHS,

$$
\min_{f \in \mathcal{H}_K} \frac{1}{N} \sum_{i=1}^{N} V(y_i, f(x_i)) + \lambda\|f\|_{\mathcal{H}_K}^2,
$$
and the optimum \( f^* \) has the following form:

\[
f^*(x) = \sum_{i=1}^{N} \alpha_i^* K(x, x_i). \tag{8.3}
\]

In large scale learning kernel methods are computationally cumbersome as they scale in the best case quadratically in the number of training examples and cubically in the worst case. Deep Neural Networks allow on the other hand rich non linear models for large scale learning and they scale linearly in the number of points, when trained with variants of stochastic gradient \([LBBH98]\).

An elegant approach pioneered by \([RR07]\), to overcome the computational load of kernel methods consists in finding explicit randomized feature maps \( \Phi : \mathcal{X} \rightarrow \mathbb{R}^m \) - where \( m \) is typically larger then the dimension of the input space \( n \) - in order to approximate the kernel \( K \) i.e.

\[
\text{For } (x, z) \in \mathcal{X}, \quad K(x, z) \approx \langle \Phi(x), \Phi(z) \rangle. \tag{8.4}
\]

Hence it is therefore sufficient to estimate a linear model using this explicit feature map \( \Phi \), and linear models scale well in large scale learning and high dimensions:

\[
\min_{w \in \mathbb{R}^m} \frac{1}{N} \sum_{i=1}^{N} V(y_i, w^\top \Phi(x_i)) + \lambda ||w||^2 \tag{8.5}
\]

Random Fourier features introduced in \([RR07]\) approximates shift invariant kernels, for example the Gaussian kernel,

\[
K(x, z) = \exp \left( -\frac{||x - z||^2}{2\sigma^2} \right),
\]

can be approximated using the following feature map:

\[
\Phi(x) = (\cos(w_1^\top x + b_1) \ldots \cos(w_m^\top x + b_m)),
\]

where \( w_i \sim \mathcal{N}(0, \frac{1}{\sigma^2} I_n) \), are independent gaussian vectors, and \( b_i \) are independently drawn uniformly from \([0, 2\pi]\).

Recently \([HAS+14]\) showed that a highly oversampled random Fourier features map, and a large scale linear least squares, matches the performance of deep neural networks on the TIMIT speech corpus.

In this chapter we focus on deep Max-out network recently introduced in \([GWFM+13]\), and try to understand their non linear abilities through introducing a Max-out random features map, and show that it leads to a locally linear estimator, and derive its equivalent kernel. In Section 8.2 we recall the architecture of deep Max-out networks, in Section 8.3 we introduce
Random Max-out features.

8.2 Deep Max-Out Networks

In Deep Max-out networks [GWFM+13] the non-linearity in the hidden units consists in taking the maximum of projections on a pool of \( q \) hyperplanes:

\[
h(x) = \max_{j=1...q} \langle w_j, x \rangle + b_j,\tag{8.6}
\]

these networks are trained with standard back propagation and stochastic gradient with dropout, and lead to state of the art results in many vision and audio classification tasks [GWFM+13, ZTPK14].

8.3 Random Max-Out Features

Random max out features have the same structure as deep max-out network in term of Max-out units (for simplicity we set all biases equal to zero and work with unit norm vectors).

The following definition gives a precise definition of the max-out random feature map:

**Definition 13** (Random Max-Out Features). Let \( w^\ell_j, \ell = 1...m, \) and \( j = 1...q, \) be independent random gaussian vectors i.e \( w^\ell_j \sim \mathcal{N}(0, I_n). \)

For \( x \in \mathbb{R}^n, \) we define a max-out random unit \( h^\ell(x) \) as follows:

\[
h^\ell(x) = \max_{j=1...q} \langle w^\ell_j, x \rangle, \quad \ell = 1...m.\tag{8.7}
\]

A max-out random feature map \( \Phi \) is therefore defined as follows:

\[
\Phi(x) = \frac{1}{\sqrt{m}}(h_1(x), ..., h_m(x)),\tag{8.8}
\]

**Remark 5.** Note that this feature map corresponds to the un-quantized \( q \)-ary sensing studied in the previous chapter where we replaced the simplex codes vectors with the canonical basis.

In order to study this map we shall study for 2 points \( x, z \in \mathbb{R}^n, \) the dot product:

\[
\langle \Phi(x), \Phi(z) \rangle = \frac{1}{m} \sum_{\ell=1}^m h^\ell(x) h^\ell(z).\tag{8.9}
\]

Consider first the expectation of \( \langle \Phi(x), \Phi(z) \rangle \):

\[
\mathbb{E}((\Phi(x), \Phi(z))) = \frac{1}{m} \sum_{\ell=1}^m \mathbb{E}(h^\ell(x) h^\ell(z)) = \mathbb{E}(h_1(x) h_1(z)),
\]

97
where the last equality follows from the independence of the units.
It is therefore sufficient to study the expectation of the dot product of one unit:

\[ K(x, z) = \mathbb{E}(h(x)h(z)), \]  

(8.10)

where \( h(x) = \max_{j=1...q} \langle w_j, x \rangle, w_j \sim \mathcal{N}(0, I_n), j = 1...q, \text{iids.} \)

**Proposition 2** (Max-Out Expected Kernel). Let \( x, z \in \mathbb{R}^n, \text{ such that } ||x|| = ||z|| = 1. \) The expected kernel of maxout random units is given by the following expression:

\[ K(x, z) = \mathbb{E}(h(x)h(z)) = \sigma^2(q) \langle x, z \rangle \kappa_q(x, z), \]  

(8.11)

where \( \sigma^2(q) = \mathbb{E}(\max_{j=1...q} g_j)^2, g_j \sim \mathcal{N}(0, 1) \text{ iid, and } \kappa_q(x, z) \text{ is a non linear kernel given by:} \)

\[ \kappa_q(x, z) = \mathbb{P}\left\{ \arg \max_{j=1...q} \langle w_j, x \rangle = \arg \max_{j=1...q} \langle w_j, z \rangle \right\} = \sum_{i=0}^{\infty} a_i(q)(\langle x, z \rangle)^i, \]

where the first 3 coefficients are \( a_0(q) = \frac{1}{q}, a_1(q) = \frac{h_2^2(q)}{q-1}, a_2(q) = \frac{q h_2^2(q)}{(q-1)(q-2)}, \) where \( h_2(q) = \mathbb{E} \phi_1(\max_{k=1...q} g_j), \) \( g_j, j = 1...q \text{ are iid standard centered gaussian, and } \phi_i, \text{ the normalized Hermite polynomials. } a_i(q) \text{ are non negative and } \sum_{i=0}^{\infty} a_i(q) = 1. \)

**Proof of Proposition 2.** The proof is given in Section 8.5. \( \square \)

### 8.3.1 Discussion of the Derived Max-Out Kernel

The expected kernel of a max-out unit is therefore a locally weighted linear kernel, and hence it allows a non linear estimation of functions in a piecewise linear way:

\[ K(x, z) = \sigma^2(q) \langle x, z \rangle \kappa_q(x, z), \]

where \( \kappa_q \) is a non linear kernel. Let \( \rho = \langle x, z \rangle. \) In this section we discuss the locality introduced by \( \kappa_q(\cdot, \cdot). \)

It is important to note that \( 0 \leq \kappa_q(x, z) \leq 1, \) since \( \kappa_q(x, z) = \mathbb{P}(D(x) = D(z)), \) where \( D(x) = \arg \max_{j=1...q} \langle w_j, x \rangle. \) Let \( x, z \in \mathbb{S}^{n-1}, \) we start first by giving values of \( \kappa_q \) in three particular cases of interest:

1. When \( x \text{ and } z \text{ coincide i.e } x = z, \text{ and } \rho = 1, \text{ we have } \kappa_q(x, z) = 1, \text{ as } \sum_{i=0}^{\infty} a_i(q) = 1. \)
2. When \( x \text{ and } z \text{ are orthogonal i.e } \rho = 0, \text{ we have } \kappa_q(x, z) = a_0(q) = \frac{1}{q}. \)
3. When \( x \text{ and } z \text{ are diametraly opposed i.e } x = -z, \text{ and } \rho = -1, \text{ we have } \kappa_q(x, z) = 0, \text{ as } \sum_{i=0}^{\infty} a_{2i}(q) = \sum_{i=0}^{\infty} a_{2i+1}(q) = \frac{1}{2}[\text{AM95}]. \)
In order to understand the locality introduced by the non linear kernel $\kappa_q$, and the relation of the radius of the locality to the size of the pool $q$, looking to the first order expansion of $\kappa_q$ gives us a hint on the effect of that kernel. In particular the quantity $h_1(q)$ is the just the expectation of maximum of independent gaussians $h_1(q) \sim \sqrt{2 \log(q)}$ \cite{Gal40, AM95}.

\[
\kappa_q(x, z) = a_0(q) + a_1(q) \rho + O(\rho^2)
\]
\[
= \frac{1}{q} (1 + (1 + \epsilon(q))2 \log(q)) + O(\rho^2),
\]

where $\epsilon(q) \to 0$, for $q \to \infty$.

Note by $g$ the function, $g : [-1, 1] \to [0, 1]$ such that $\kappa_q(x, z) = g(\rho)$. For far apart points, when $\rho \to -1$, $g(\rho) \to 0$. $g$ has a linear behavior around $\rho = 0$, with a slope equal to $\frac{2 \log(q)}{q}$. Note that in this neighborhood as $q$ increases the linear regime vanishes, and $g(\rho) \to 0$. Hence as $q$ increases the probability of two points hashing to same index of maximum becomes smaller, qualitatively the radius of the locality of $\kappa_q$ shrinks as $q$ increases. Finally for near by points when $\rho \to 1$, $g(\rho) \to 1$.

Qualitatively the derived kernel $K(x, z) \approx 0$ for far apart points and $K(x, z) = \sigma^2(q) \langle x, z \rangle$ for points in the same neighborhood, where the radius of the locality, and the notion of closeness is set by the choice of the size of the pool $q$. This radius is decreasing in $q$. Hence $K$ defines a locally linear kernel.

Now if we go back to problem (8.2), and solve for $f$ in the reproducing kernel hilbert space
of the equivalent kernel $K$ (i.e for $\mathcal{H} = \mathcal{H}_K$), we have:

$$f^*(x) = \sum_{i=1}^{N} \alpha_i^* K(x, x_i) = \sigma^2(q) \sum_{i=1}^{N} \alpha_i^* \langle x, x_i \rangle \kappa_q(x, x_i)$$  \hspace{1cm} (8.12)

$$= \sigma^2(q) \left\langle \sum_{i=1}^{N} \alpha_i^* \kappa_q(x, x_i) x_i, x \right\rangle .$$  \hspace{1cm} (8.13)

Hence we see that this derived kernel allows a locally linear estimation of the function of interest $f^*$, where the radius of the locality is set by the choice of the size of the pool $q$.

Now consider the max-out random feature map $\Phi$ introduced in Definition 13 recall that we have:

$$\mathbb{E}(\langle \Phi(x), \Phi(z) \rangle) = K(x, z) = \sigma^2(q) \langle x, z \rangle \kappa_q(x, z),$$

the dot product $\langle \Phi(x), \Phi(z) \rangle$ is therefore an estimator of $K(x, z)$ i.e for sufficiently large $m$, $\langle \Phi(x), \Phi(z) \rangle \approx K(x, z)$, hence we can use the feature map $\Phi$, and a simple linear model as in equation (8.5), and use the optimal weight $w^*$ to get an estimate of the locally linear estimation $f^*$ produced by the derived kernel as in equation (8.13), i.e. we have for sufficiently large $m$:

$$\langle w^*, \Phi(x) \rangle \approx f^*(x).$$  \hspace{1cm} (8.14)

### 8.3.2 Locality Sensitive Hashing

Let $C : \mathbb{S}^{n-1} \rightarrow \{1 \ldots q\}^m$, such that for $x \in \mathbb{S}^{n-1}$

$$C(x) = \left( \arg \max_{j=1 \ldots q} \langle w_j^1, x \rangle, \ldots, \arg \max_{j=1 \ldots q} \langle w_j^m, x \rangle \right), \quad w_j^\ell \sim \mathcal{N}(0, I_n), \ell = 1 \ldots m, j = 1 \ldots q.$$

For $x, z \in \mathbb{S}^{n-1}$ we have:

$$\mathbb{E} \left( \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{C_i(x) \neq C_i(z)} \right) = P \left\{ \arg \max_{j=1 \ldots q} \langle w_j, x \rangle \neq \arg \max_{j=1 \ldots q} \langle w_j, x \rangle \right\} = 1 - \kappa_q(x, z).$$

Hence we can approximate the local kernel $\kappa_q(x, z)$, by the non binary strings by mean of the hamming distance between the $q$-ary strings $C(x)$, and $C(z)$. As $m$ becomes large we have:

$$d_H(C(x), C(z)) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{C_i(x) \neq C_i(z)} \approx 1 - \kappa_q(x, z),$$

Hence $C$ defines a locality sensitive hashing scheme in the sense of [Ind01].

100
8.3.3 Sample Complexity of the Random Max-Out Feature Map

We assume that the points we are interested in embedding through the maxout random feature map $\Phi$, are on the unit sphere $S^{n-1}$, and belong to a set $K$. We note $\mathcal{M} = K \cap S^{n-1}$. We wish to have bound on the following quantity:

$$\Delta = \frac{1}{\log(q)} \sup_{x,z \in \mathcal{M}} \left| K(x, z) - \langle \Phi(x), \Phi(z) \rangle \right|, \quad (8.15)$$

Note that we normalized by $\frac{1}{\log(q)} \sim \sigma^2(q)$, so that the kernel we are approximating is $\langle x, z \rangle \kappa_q(x, z)$, and the feature map is $\frac{1}{\sqrt{\log(q)}} \Phi(x)$.

We will assume in the following a notion of intrinsic dimension on the set $\mathcal{M}$, namely the Assouad dimension given in the following definition:

**Definition 14** ([Hei01]). The Assouad dimension of $\mathcal{M} \subset \mathbb{R}^n$, denoted by $d_{\mathcal{M}}$, is the smallest integer $k$, such that, for any ball $B \subset \mathbb{R}^n$, the set $B \cap \mathcal{M}$ can be covered by $2^k$ balls of half the radius of $B$.

The Assouad dimension is a widely used measure of the intrinsic dimension. For example, if $\mathcal{M}$ is an $\ell_p$ ball in $\mathbb{R}^n$, then $d_{\mathcal{M}} = O(n)$ if $\mathcal{M}$ is a $d$-dimensional hyperplane in $\mathbb{R}^n$, then $d_{\mathcal{M}} = O(d)$. Moreover, if $\mathcal{M}$ is a $d$-dimensional Riemannian submanifold of $\mathbb{R}^n$ with a suitably bounded curvature, then $d_{\mathcal{M}} = O(d)$.

**Proposition 3** (Approximation Error in Expectation). For $\epsilon > 0$, the following holds:

$$m \geq Cn \frac{d_{\mathcal{M}}}{\epsilon^2 \log(q)} \frac{1}{\log(q)} \mathbb{E} \sup_{x,z \in \mathcal{M}} \left| \langle \Phi(x), \Phi(z) \rangle - K(x, z) \right| \leq \epsilon$$

**Proof of Proposition 3.** The proof is given in Section 8.5. \qed

From Proposition 3, we see that the dimension of the embedding $m$ is bigger than the original dimension $n$, $m > n$, and the oversampling factors depends on the intrinsic dimension $d_{\mathcal{M}}$. If there was no structure, we need $O(n^2)$ dimensions, with some structure say that $\mathcal{M}$ is a manifold of dimension $r$ in $\mathbb{R}^n$, than the oversampling factor is proportional to $r$, $m = O(rn)$. The bounds we get here are not sharp, and this is due to the technique of the proof, better techniques would lead to tighter bounds and a better dependency on $n$, and $d_{\mathcal{M}}$. We conjecture that the optimal embedding is of dimension $O(d_{\mathcal{M}} \log(n))$.

8.4 Numerical Experiments

We extracted the Maxout random features on the MNIST dataset, in that case the dimension $n = 784$. We than run a stochastic gradient descent on the cross-entropy loss, the cross
validation was done using hold out to set the number of epochs. We consider \( m = 15000 \), and various pool size \( q \), error rates are summarized in the following table:

<table>
<thead>
<tr>
<th>( q )</th>
<th>error</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>2.08%</td>
</tr>
<tr>
<td>8</td>
<td>1.89%</td>
</tr>
<tr>
<td>16</td>
<td>1.80%</td>
</tr>
<tr>
<td>32</td>
<td>1.86%</td>
</tr>
<tr>
<td>Averaged posteriors</td>
<td>1.74%</td>
</tr>
</tbody>
</table>

Table 8.1: Random Maxout on MNIST.

A linear classifier on the pixels achieves 15\% error rates. Maxout neural networks with no data augmentation achieve higher performance 0.7\% error rate. The error of a fine tuned RBF kernel SVM is of 1.4\%, the error of a nearest neighbors is worst 3.09\%. Random Max-Out feature maps with a linear model achieves an error rate slightly worst than kernel methods but better than nearest neighbors.

8.5 Proofs of Propositions 2 and 3

**Proof of Proposition 2.** Assume that \( ||x|| = ||z|| = 1 \). Let \( D(x) = \arg \max_{j=1...q} \langle w_j, x \rangle \), and \( D(z) = \arg \max_{j=1...q} \langle w_j, z \rangle \), ties are broken arbitrarily. By total probability we have:

\[
\mathbb{E}(h(x)h(z)) = \mathbb{E}\{h(x)h(z)|D(x) = D(z)\} \mathbb{P}(D(x) = D(z)) \\
+ \mathbb{E}\{h(x)h(z)|D(x) \neq D(z)\} \mathbb{P}(D(x) \neq D(z)) \\
= \mathbb{E}(\langle w_{D(x)}, x \rangle \langle w_{D(x)}, z \rangle |D(x) = D(z)) \mathbb{P}(D(x) = D(z)) \\
+ \mathbb{E}\{\langle w_{D(x)}, x \rangle \langle w_{D(z)}, z \rangle |D(x) \neq D(z)\} \mathbb{P}(D(x) \neq D(z)).
\]

It is easy to see that the second term in this sum is zero since the gaussians are independent and zero centered: \( \mathbb{E}\{\langle w_{D(x)}, x \rangle \langle w_{D(z)}, z \rangle |D(x) \neq D(z)\} = 0 \). We are left with the first term of this sum:

\[
\mathbb{E}(h(x)h(z)) = \mathbb{E}(\langle w_{D(x)}, x \rangle \langle w_{D(x)}, z \rangle |D(x) = D(z)) \mathbb{P}(D(x) = D(z)) \\
= q\mathbb{E}(\langle w_1, x \rangle \langle w_1, z \rangle |D(x) = D(z) = 1) \mathbb{P}(D(x) = D(z) = 1) \\
= \mathbb{E}(\langle w_1, x \rangle \langle w_1, z \rangle |D(x) = D(z) = 1) \mathbb{P}(D(x) = D(z)).
\]

By rotation invariance of gaussians we have:

\[
\langle w_1, x \rangle = g \text{ and } \langle w_1, z \rangle = \langle x, z \rangle g + \sqrt{1 - ||x, z||^2} h,
\]

102
where \( g \) and \( h \) are independent random gaussian variables \( g, h \sim \mathcal{N}(0, 1) \).

Let \( E \) be the following event:

\[
E = \{ g \text{ is the maximum of } q \text{ independent gaussians} \}
\]

Hence we have:

\[
\mathbb{E}(\langle w_1, x \rangle \langle w_1, z \rangle | D(x) = D(z) = 1) = \mathbb{E} \left( g(\langle x, z \rangle) g + \sqrt{1 - |\langle x, z \rangle|^2} h | E \right) = \langle x, z \rangle \mathbb{E} \left( \left[ \max_{j=1...q} g_j \right]^2 \right).
\]

Let \( \sigma^2(q) = \mathbb{E} \left( [\max_{j=1...q} g_j]^2 \right) \), we have finally:

\[
\mathbb{E}(h(x)h(z)) = \sigma^2(q) \langle x, z \rangle \mathbb{P}(D(x) = D(z)). \tag{8.16}
\]

\( \sigma^2(q) \) is a normalization factor and it is well known that \( \sigma^2(q) \sim \log(q) \), hence we are left with

\[
\mathbb{P}(D(x) = D(z)),
\]

that is the probability that \( x \) and \( z \) are not separated by the \( q \) hyperplanes, an object that is well studied in \( q \)-ways graph cuts approximation algorithm.

The following lemma is crucial to our proof and is proved in \([AM95]\), and allow us to get the final expression of the expected kernel.

**Lemma 8** ([AM95]). For \( x, z \in \mathbb{R}^n, ||x|| = ||z|| = 1 \). Let \( \rho = \langle x, z \rangle \), we have therefore:

\[
\kappa_q(x, z) = \mathbb{P}(D(x) = D(z)) = \sum_{i=0}^{\infty} a_i(q) \rho^i = a_0(q) + a_1(q) \rho + a_2(q) \rho^2 + \ldots \tag{8.17}
\]

the taylor series of \( \kappa_q \) around \( \rho = 0 \), converges for all \( \rho \) in the range \( |\rho| \leq 1 \). The coefficients \( a_i(q) \), of the expansion are all non negatives and their sum converges to 1. The first 3 coefficients are \( a_0(q) = \frac{1}{q}, a_1(q) = \frac{h_2^2(q)}{q-1}, a_2(q) = \frac{q h_3(q)}{(q-1)(q-2)} \). \( h_i(q) = \mathbb{E} \phi_i(\max_{k=j...q} \eta_j) \), where \( \eta_j, j = 1 \ldots q \) are iid standard centered gaussian, and \( \phi_i \), the normalized Hermite polynomials.

By lemma 8 we have finally:

\[
K(x, z) = \mathbb{E}(h(x)h(z)) = \sigma^2(q) \langle x, z \rangle \kappa_q(x, z), \tag{8.18}
\]

where \( \kappa_q(x, z) = \sum_{i=0}^{\infty} a_i(q) |\langle x, z \rangle|^i \), is a non linear kernel, values of \( a_i(q) \) are given in the above lemma.

Before stating the proof of Proposition 3, we introduce in the following section some
useful technical tools and definitions needed for the proof.

### 8.5.1 Technical tools: Covering and Bracketing Numbers

Let \((\mathcal{F}, ||\cdot||)\), be a metric space. We start with the following preliminary definitions, of covering and bracketing numbers.

**Definition 15** (Covering Numbers,[vdVW96]). The covering number \(\mathcal{N}(\epsilon, \mathcal{F}, ||\cdot||)\) is the minimal number of balls of radius \(\epsilon\), needed to cover the set \(\mathcal{F}\). The centers of the balls need not to belong to \(\mathcal{F}\), but they should have finite norms. The entropy is the logarithm of the covering number.

**Definition 16** (Bracketing Numbers,[vdVW96]). Given two functions \(l\), and \(u\), the bracket \([l, u]\) is the set of all functions \(f\) with \(l \leq f \leq u\). An \(\epsilon\)-bracket is a bracket with \(||u - l|| < \epsilon\). The bracketing numbers \(\mathcal{N}_{[\cdot]}(\epsilon, \mathcal{F}, ||\cdot||)\) is the minimum number of \(\epsilon\)-brackets needed to cover \(\mathcal{F}\). The entropy with bracketing is the logarithm of the bracketing number. In the definition of the bracketing number, the upper and lower bounds \(u\) and \(l\) of the brackets need not to belong to \(\mathcal{F}\) themselves but are assumed to have finite norms.

The following lemma relates covering numbers to bracketing numbers and will be useful in our derivations:

**Lemma 9** ([vdVW96]). Let \(\mathcal{N}_{[\cdot]}(\delta, \mathcal{F}, ||\cdot||_{L^2(\mu)})\) be the bracketing number of the class \(\mathcal{F}\), with respect to the norm \(||\cdot||_{L^2(\mu)}\), i.e for \(f \in \mathcal{F}\), \(f\) has the following form \(f(W)\) where \(W\) is distributed according to \(\mu\), we have:

1. The covering numbers are bounded by the bracketing numbers:

\[
\mathcal{N}(\delta, \mathcal{F}, ||\cdot||_{L^2(\mu)}) \leq \mathcal{N}_{[\cdot]}(2\delta, \mathcal{F}, ||\cdot||_{L^2(\mu)}).
\] (8.19)

2. Let \(\mathcal{F}\), and \(\mathcal{G}\) be two classes, bounded by 1, we have:

\[
\mathcal{N}_{[\cdot]}(2\delta, \mathcal{F} \cup \mathcal{G}, ||\cdot||_{L^2(\mu)}) \leq \mathcal{N}_{[\cdot]}(\delta, \mathcal{F}, ||\cdot||_{L^2(\mu)}) \mathcal{N}_{[\cdot]}(\delta, \mathcal{G}, ||\cdot||_{L^2(\mu)}).
\] (8.20)

3. \(\mathcal{F} = \{f : W \to f_x(W)\}, x \in \mathcal{M}\), a class of functions indexed by a set \(\mathcal{M}\) such that

\[
|f_x(W) - f_z(W)| \leq d(x, z)F(W), \forall W,
\] (8.21)

where \(d\) is a distance defined on \(\mathcal{M}\). then we have:

\[
\mathcal{N}_{[\cdot]}(2\epsilon \cdot ||F||_{L^2(\mu)}, \mathcal{F}, ||\cdot||_{L^2(\mu)}) \leq \mathcal{N}(\epsilon, \mathcal{M}, d).
\] (8.22)
Proof of Proposition 3. We assume that the points we are interested in embedding through the maxout random feature map \( \Phi \), are on the unit sphere \( S^{n-1} \), and belong to a set \( K \). We note \( \mathcal{M} = K \cap S^{n-1} \). We wish to have bound on the following quantity:

\[
\Delta(W) = \sup_{x,z \in \mathcal{M}} |K(x, z) - \langle \Phi(x), \Phi(z) \rangle|.
\]  

Define

\[
\mathcal{F} = \left\{ f_{x}(w_1, \ldots, w_q) = \max_{j=1, \ldots, q} \langle w_j, x \rangle, x \in \mathcal{M} \right\},
\]

and let \( W_\ell = (w_{1\ell}, \ldots, w_{q\ell}) \), \( \ell = 1 \ldots m \), and \( W = (w_1 \ldots w_q) \), we have therefore

\[
\Delta(W) = \sup_{x \in \mathcal{M}, z \in \mathcal{M}} \left| \frac{1}{m} \sum_{\ell=1}^{m} f_{x}(W_\ell) f_{z}(W_\ell) - \mathbb{E}(f_{x}(W) f_{z}(W)) \right|.
\]

By symmetrization we have:

\[
\mathbb{E}(\Delta(W)) \leq 2 \mathbb{E}_{\epsilon,W} \left( \sup_{x, z \in \mathcal{M}} \frac{1}{m} \left| \sum_{\ell=1}^{m} \epsilon_{\ell} \max_{j=1, \ldots, q} \langle w_{j\ell}, x \rangle \max_{j=1, \ldots, q} \langle w_{j\ell}, z \rangle \right| \right),
\]

where \( \epsilon_{\ell} \) are iid, rademacher random variables \( \mathbb{P}(\epsilon_{\ell} = 1) = \mathbb{P}(\epsilon_{\ell} = -1) = \frac{1}{2} \).

\[
\mathcal{R}_m(\mathcal{F}) = \mathbb{E}_{\epsilon,W} \left( \sup_{x, z \in \mathcal{M}} \frac{1}{m} \left| \sum_{\ell=1}^{m} \epsilon_{\ell} f_{x}(W_\ell) f_{z}(W_\ell) \right| \right).
\]

The quantity \( \mathcal{R}_m(\mathcal{F}) \) can be bounded by the dudley entropy integral:

\[
\mathcal{R}_m(\mathcal{F}) \leq 2 \frac{C_0}{\sqrt{m}} \int_{0}^{+\infty} \sqrt{\log(\mathcal{N}(\epsilon, \mathcal{F}, ||.||_{L^2}))} d\epsilon,
\]

where \( C_0 > 0 \) is a universal constant, \( \mathcal{N}(\epsilon, \mathcal{F}, ||.||_{L^2}) \) is the \( \epsilon \) covering number of the piecewise linear function class \( \mathcal{F} \), with respect to the \( L^2(\mu) \) norm, where \( \mu \) is the distribution of \( (w_1 \ldots w_q), (w_j \sim \mathcal{N}(0, I_n), j = 1 \ldots q \) iid). We are left now with bounding these covering numbers of \( \mathcal{F} \), with the covering numbers of \( \mathcal{M} \) with respect to the euclidean norm in \( \mathbb{R}^n \) (\( \mathcal{N}(\delta, \mathcal{M}, ||.||_2) \)).

Assume we have: \( |f_{x}(W) - f_{z}(W)| = \langle w_{D(x)}, x \rangle - \langle w_{D(z)}, z \rangle \), meaning \( \langle w_{D(x)}, x \rangle - \langle w_{D(z)}, z \rangle > 0 \).

\[
f_{x}(W) - f_{z}(W) = \langle w_{D(x)}, x \rangle - \langle w_{D(z)}, z \rangle = \langle W_{D(x)}, x - z \rangle - \langle w_{D(z)} - w_{D(x)}, z \rangle \geq 0 \text{ By definition of } D(z).
\]

\[
\leq ||w_{D(x)}||_2 ||x - z||_2 \text{ By Cauchy-Schawrz inequality.}
\]
Similarly if we have \(|f_x(W) - f_z(W)| = \langle w_{D(x)}, z \rangle - \langle w_{D(x)}, x \rangle\), we have:

\[
f_x(W) - f_z(W) \leq \|w_{D(z)}\|_2\|x - z\|_2
\]

We conclude therefore that:

\[
|f_x(W) - f_z(W)| \leq \max (\|w_{D(x)}\|_2, \|w_{D(z)}\|_2) \|x - z\|_2.
\]  

Let \(L = 2\|w\|_{L^2(\mu)} = 2\sqrt{n}\), and \(\sigma(q) \sim \sqrt{2\log(q)}\). We have by Lemma 9:

\[
\mathcal{N}(\delta, F, F, ||\cdot||_{L^2(\mu)}) \leq \mathcal{N}(2\delta, F, F, ||\cdot||_{L^2(\mu)}) \leq \left(\mathcal{N}_{\delta} \left(\delta, F, F, ||\cdot||_{L^2(\mu)}\right)\right)^2 \leq \left(\mathcal{N} \left(\frac{\delta}{2L\sigma(q)}, M, ||\cdot||_2\right)\right)^2
\]

Let \(d_M\) be the Assouad dimension of \(M\), and \(\text{diam}(M)\) be the diameter of \(M\), we have by definition of the Assouad dimension:

\[
\mathcal{N}(\delta, M, ||\cdot||_2) \leq \left(\frac{2\text{diam}(M)}{\delta}\right)^{d_M},
\]  

then we have

\[
\mathcal{N}(\delta, F, F, ||\cdot||_{L^2(\mu)}) \leq \left(\frac{\text{diam}(M)L\sigma(q)}{\delta}\right)^{2d_M}
\]

Hence:

\[
\log (\mathcal{N}(\delta, F, F, ||\cdot||_{L^2(\mu)})) \leq 2d_M \log \left(\frac{\text{diam}(M)L\sigma(q)}{\delta}\right),
\]

Hence

\[
R_m(F) \leq 2C_0 \sqrt{\frac{2d_M}{m}} \int_0^{+\infty} \sqrt{\log \left(\frac{\text{diam}(M)L\sigma(q)}{\delta}\right)} d\delta
\]

\[
\leq C_1 \sqrt{\frac{d_M}{m} \text{diam}(M) \sqrt{n \log(q)}}.
\]

Hence:

\[
\frac{1}{\log(q)} R_m(F) \leq C_1 \sqrt{\frac{nd_M}{m \log(q)}} \text{diam}(M).
\]  

(8.27)

Since we are on the sphere \(\text{diam}(M) \leq 2\) hence for

\[
m \geq Cn \frac{d_M}{\varepsilon^2 \log(q)} \frac{1}{\log(q)} \sup_{x, z \in M} |\Phi(x) - \Phi(z)| \leq \varepsilon.
\]

the oversampling factors depends on the intrinsic dimension \(d_M\). If there was no structure, we have therefore \(O(n^2)\), with some structure say that \(M\) is a manifold of dimension \(r\) in \(\mathbb{R}^n\), than the oversampling factor is proportional to \(r\), \(m = O(rn)\).
Part III

Phase Retrieval
Chapter 9

One Bit Phase Retrieval

In this chapter, we study the problem of robust phase recovery. We introduce and investigate a novel approach based on extremely quantized (one-bit) measurements and a corresponding recovery scheme. The proposed approach has surprising robustness properties and, unlike currently available methods, allows to efficiently perform phase recovery from measurements affected by severe (possibly unknown) non linear perturbations, such as distortions (e.g. clipping). Beyond robustness, we show how our approach can be used within greedy approaches based on alternating minimization. In particular, we propose novel initialization schemes for the alternating minimization achieving favorable convergence properties with improved sample complexity.

9.1 Introduction

The phase recovery problem can be modeled as the problem of reconstructing a $n$-dimensional complex vector $x_0$ given only the magnitude of $m$ phase-less linear measurements. Such a problem arises for example in X-ray crystallography [Har93, Lea08], diffraction imaging [BDP+07, Rod08] or microscopy [MISE08], where one can only measure the intensities of the incoming waves, and wishes to recover the lost phase in order to be able to reconstruct the desired object.

In practice, phase recovery is often tackled via greedy algorithms [GS72, Fie82, GL84] which typically lack convergence guarantees. Recently, approaches based on convex relaxations, namely PhaseLift in [CSV11, DH12], and Phase cut in [WDM12], have been proposed and analyzed. These latter methods can be solved by Semi Definite Programing (SDP), and allow the exact and stable recovery of the signal (up to a global phase) from $O(n)$ measurements. A different approach have been recently considered in [NJS13], where it is shown that a greedy alternating minimization, akin to those in [GS72, Fie82, GL84], can be shown to geometrically converge to the true vector $x_0$ if $O(n \log^3(n))$ measurements are given. Indeed, alternating minimization algorithms are known to be extremely sensitive to the initialization and a suitable initialization is the key of the analysis in [NJS13]. Another non convex
approach for phase retrieval based on gradient descent, namely Wirtinger flow was recently introduced in [CLS14], and allows for approximate recovery from $O(n \log n)$ Gaussian measurements.

While non convex approaches provide a solution only up-to a given accuracy, they often have very good practical performances when compared to convex methods [NJS13], with dramatic computational advantages [NJS13]. The solution of the SDP in convex approaches is computationally expensive and needs to be close to a rank one matrix for tight recovery (which is rarely encountered in practice [WDM12]). Indeed, some greedy refinement of the SDP solution is often considered [WDM12].

In this chapter, we propose and investigate a phase recovery approach based on extremely quantized measurements and a corresponding recovery procedure. In particular, we study the properties of the proposed method towards the following questions:

1. **Robustness**: Is it possible to efficiently perform phase recovery, when the measurements are corrupted by sever perturbations such as non linear distortions or stochastic noise?

2. **Refinements of Alternating Minimization**: Are there initialization strategies for the alternating minimization approach that allow better sample complexity?

Robustness to noise and distortions, such as clipping of the intensities or imperfections in Fourier optics such as multiple scattering [MK83] is a desirable property for a phase retrieval algorithm. At first this task might seem hopeless since current approaches to phase recovery are based on measurements magnitudes which might be completely altered by distortions or if the signal-to-noise ratio is very poor. In fact, we prove the somewhat surprising fact that phase retrieval is still possible, as long as the perturbations preserve (on average) the ranking of the measurements intensities. Indeed, key to our approach is considering a suitable quantization scheme based on comparing pairs of phase-less measurements: only the ranking of each measurement pairs becomes important, rather than the intensity values themselves. Using these extremely quantized (one-bit) measurements, we show that recovery is possible as soon as $O(n)$ pairs of measurements are available. The corresponding recovery procedure reduces to a maximum eigenvalue problem (1bitPhase) which can be efficiently solved, for example using the power method. Our approach is inspired by the growing field of one-bit compressive sensing [BB08, PV13b, PV13a, JLBB13].

Beyond robustness, we show how the nature of the one-bit phase less measurements can be used to obtain better results for alternating minimization. We show that the solution of one-bit phase retrieval can be used to initialize alternating minimization to obtain the same convergence results in [NJS13] from only $O(n)$ measurements.

The rest of the chapter is organized as follows. In Section 9.2, we discuss some background and previous results. In Section 9.3, we sketch our main results and techniques. In Section 9.4, we introduce and analyze the One-Bit Phase Retrieval approach. We provide a theoretical analysis of our approach in Section 9.5. In Section 9.6 we discuss how one-bit
phase retrieval can be used to initialize the alternating minimization approach to get a better sample complexity. Finally, in Section 10.5 we discuss some computational aspects and present some numerical results.

Notations: For \( z \in \mathbb{C} \), \( |z|^2 \) is squared complex modulus of \( z \). For \( a, a' \in \mathbb{C}^n \), \( \langle a, a' \rangle \) is the complex dot product in \( \mathbb{C}^n \). For \( a \in \mathbb{C}^n \), \( a^* \) is the complex conjugate and \( ||a||_2 \) or simply \( ||a|| \) is the norm 2 of \( a \). Let \( A \) a complex hermitian matrix in \( \mathbb{C}^n \), \( ||A||_F \) denotes the Frobenius norm of \( A \), \( ||A|| \) denotes the operator norm of \( A \), \( Tr(A) \) denotes the trace of \( A \). Throughout the chapter, we denote by \( c, C \) positive absolute constants whose values may change from instance to instance.

9.2 Background and Previous Work

In this section, we formalize the problem of recovering a signal from phase-less measurements and discuss previous results. Throughout this section, and the rest of the chapter, we consider measurements defined by independent and identically distributed Complex Gaussian sensing vectors,

\[ a_i \in \mathbb{C}^n, \quad a_i \sim \mathcal{N}(0, \frac{1}{2}I_n) + \mathcal{N}(0, \frac{1}{2}I_n), \quad i = 1 \ldots m. \tag{9.1} \]

The (noiseless) phase recovery problem is defined as follows.

**Definition 17 (Phase-less Sensing and Phase Recovery).** Suppose phase-less sensing measurements

\[ b_i = |\langle a_i, x_0 \rangle|^2 \in \mathbb{R}_+, \quad i = 1 \ldots m, \tag{9.2} \]

are given for \( x_0 \in \mathbb{C}^n \), where \( a_i, i = 1, \ldots, m \) are random vectors as in (9.1). The phase recovery problem is

\[ \text{find } x, \quad \text{subject to } |\langle a_i, x \rangle|^2 = b_i, \quad i = 1 \ldots m. \tag{9.3} \]

The above problem is non convex and in the following we recall recent approaches to provably and efficiently recover \( x_0 \) from a finite number of measurements.

**SDP (Convex) Relaxation and PhaseLift.** The PhaseLift approach [CSV11] stems from the observation that \( |\langle a_i, x \rangle|^2 = Tr(a_i a_i^* xx^*), \) so that if we let \( X = xx^* \), Problem 9.3 can be written as,

\[ \text{find } X, \quad \text{subject to } Tr(a_i a_i^* X) = b_i, \quad i = 1 \ldots m, \quad X \succeq 0, \quad \text{rank}(X) = 1. \tag{9.4} \]

While the above formulation is still non convex (and in fact combinatorially hard because of the rank constraint), a convex relaxation can be obtained noting that Problem 9.4 can be
written as a rank minimization problem over the positive semidefinite cone,

\[
\min_X \text{rank}(X), \quad \text{subject to} \quad Tr(a_i a_i^* X) = b_i, \quad i = 1 \ldots m, \quad X \succeq 0, \quad (9.5)
\]

and then considering the trace as a surrogate for the rank [CSV11],

\[
\min_X \text{Tr}(X), \quad \text{subject to} \quad Tr(a_i a_i^* X) = b_i, \quad i = 1 \ldots m, \quad X \succeq 0. \quad (9.6)
\]

Indeed, the above problem is convex and can be solved via semidefinite programming (SDP).

Interestingly, a different relaxation is obtained in [DH12] by ignoring the rank constraint in Problem 9.4. The results in [CSV11, DH12] show that, with high probability, the solution \( \hat{X}_m \) obtained via either one of the above relaxations can recover \( x_0 \) exactly, i.e. \( \hat{X}_m = x_0 x_0^* \), as soon as \( m \geq cn \log n \). In fact, the latter requirement can be further improved to \( m \geq cn \) [CL12]. If the measurements are corrupted by noise, namely

\[
b_i = |\langle a_i, x_0 \rangle|^2 + w_i, \quad i = 1 \ldots m, \quad (9.7)
\]

the PhaseLift approach can be adapted [CL12] by considering

\[
\min_X \sum_{i=1}^{m} |Tr(a_i a_i^* X) - b_i|, \quad \text{subject to} \quad X \succeq 0. \quad (9.8)
\]

The above problem is convex and can again be solved via an SDP approach. The properties of its solution have been studied in [CL12] for deterministic noise

\[
||w||_1 \leq \delta, \quad w = (w_1, \ldots, w_m) \in \mathbb{R}_+^m,
\]

where it is shown that the solution \( \hat{X}_m \) of (9.8) satisfies \( ||\hat{X}_m - x_0 x_0^*||_F \leq c\delta / m \), as soon as \( m \geq cn \). Moreover, the leading eigenvector \( \hat{x}_m \) of \( \hat{X}_m \) satisfies

\[
||\hat{x}_m - e^{i\phi} x_0||_2 \leq c \min \left( ||x_0||_2, \frac{\delta}{m||x_0||_2} \right),
\]

where \( \phi \) is a global phase in \([0, 2\pi]\). Most importantly, the latter results suggests that \( x_0 \) can be recovered considering the leading eigenvector of \( \hat{X}_m \).

As mentioned in the introduction, while powerful, the convex relaxation approach incur in cumbersome computations– see Table 9.1, and in practice non convex approaches based on greedy alternating minimization (AM) [GS72, Fie82, GL84] are often used. The convergence properties of the latter methods depend heavily on the initialization and only recently [NJS13] they have been shown to globally converge (with high probability) if provided with a suitable initialization.

\textbf{Phase Retrieval via suitably initialized Alternating Minimization and Resam-}
pling. Let $A$ be the matrix defined by $m$ sensing vectors as in (9.1) and $B = \text{Diag}(\sqrt{b})$, where $b$ is the vector of measurements as in (9.2). Then, $Ax_0 = Bu_0$, for $u_0 = Ph(Ax_0)$ with $Ph(z) = \left(\frac{z_1}{|z_1|}, \ldots, \frac{z_m}{|z_m|}\right)$, $z \in \mathbb{C}^m$. The above equality suggests the following natural approach to recover $(x_0, u_0)$,

$$\min_{x,u} ||Ax - Bu||^2, \quad \text{subject to} \quad |u_i| = 1, \quad i = 1 \ldots m,$$

The above problem is non-convex because of the constraint on $u$. The AM approach consists in optimizing $u$, for a given $x$, and then optimizing $x$ for a given $u$. It is easy to see that for a given $x$, the optimal $u$ is simply $u = Ph(Ax)$, and for a given $u$, the optimal $x$ is the solution of a least squares problem. The key result in [NJS13] shows that if such an iteration is initialized with maximum eigenvector of the matrix

$$\hat{C}_m = \frac{1}{m} \sum_{i=1}^{m} b_i a_i a_i^*, \quad (9.9)$$

then the solution $x_{t_0}$ of the alternating minimization (Algorithm 3) globally converges (with high probability) to the true vector $x_0$. For a given accuracy $\epsilon$, $0 < \epsilon < 1$, if

$$m \geq c n \left(\log^3 n + \log \frac{1}{\epsilon} \log \log \frac{1}{\epsilon}\right), \quad \text{then} \quad ||x_{t_0} - e^{i\phi}x_0||_2 \leq \epsilon, \quad (9.10)$$

where $\phi$ is a global phase.

A key observation, motivating the above initialization (called $\text{SubExpPhase}$), is the fact that the expectation of $\hat{C}_m$ can be shown to satisfy

$$\mathbb{E}(\hat{C}_m) = x_0 x_0^* + ||x_0||^2 I. \quad (9.11)$$

Indeed, the proof in [NJS13] relies on the concentration properties of the random matrix $\hat{C}_m$ around its expectation [Ver11, Tro12]. It is useful to note that these latter results crucially depend on a bound on the norm of $b_i a_i a_i^*$ for $i = 1, \ldots, m$. Indeed, it is this latter bound the main cause of the poly-logarithmic term in the sample complexity of $\text{SubExpPhase}$ since the $b_i$'s are exponential random variables.

Algorithm 3 proposed in [NJS13], proceeds in alternating the estimation of the phase and the signal. For technical reasons - mainly ensuring independence - the algorithm proceeds in a stage-wise alternating minimization. At each stage we use a new re-sampled sensing matrix and the corresponding measurements.
Algorithm 3 AltMinPhase with Resampling

1: procedure ALTMINPHASERESAMPLING(A, b, c)
2: \[ t_0 \leftarrow c \log \left( \frac{1}{\epsilon} \right) \]
3: Partition \( b \) and the corresponding rows of \( A \) into \( t_0 + 1 \) disjoint sets of same cardinality: 
   \((b_0, A_0), \ldots (b_{t_0}, A_{t_0})\).
4: Initialize \( x \)
5: for \( t = 0 \) to \( t_0 - 1 \) do
6: \[ u_{t+1} \leftarrow Ph(A_{t+1} \hat{x}_t) \]
7: \[ \hat{x}_t \leftarrow \arg \min \| A_{t+1} x - B_{t+1} u_{t+1} \|^2 \]
8: end for
9: return \( x_{t_0} \)
10: end procedure

Phase Retrieval via Wirtinger Flow. Consider the least squares misfit problem between the quadratic measurements of a point \( x \in \mathbb{C}^n \) and the true measurements \( \{ b_i = | \langle a_i, x_0 \rangle |^2 \} \)

\[
\min_{x \in \mathbb{C}^n} f(x) := \frac{1}{2m} \sum_{i=1}^{m} (| \langle a_i, x \rangle |^2 - b_i)^2.
\]  

(9.12)

Clearly, the function \( f \) is non convex in \( x \). Recently [CLS14] introduced a new approach to provably solve this non convex problem via suitably initialized, gradient descent based on Wirtinger derivatives. We summarize in the following the approach taken in [CLS14]:

1. The initialization step is also a spectral initialization similar to the one in [NJS13]: \( z_0 \) is set to be \( \hat{x}_m \) the maximum eigenvector of the matrix \( \hat{C}_m \) defined in Equation (9.9), normalized to \( \| z_0 \| = \sqrt{\frac{1}{m} \sum_{i=1}^{m} b_i} \).

2. Starting with the initialization \( z_0 \), the updates are given for \( \tau = 0, 1, 2 \ldots \) by:

\[
z_{\tau+1} = z_\tau - \frac{\mu_{\tau+1}}{\|z_\tau\|^2} \nabla f(z_\tau),
\]  

(9.13)

where

\[
\nabla f(z_\tau) = \frac{1}{m} \sum_{i=1}^{m} (| \langle a_i, z \rangle |^2 - b_i) a_i a_i^* z,
\]  

(9.14)

and \( \mu_{\tau+1} \) is the step size.

For \( x, x_0 \in \mathbb{C}^n \) define:

\[
\text{dist}(x, x_0) = \min_{\phi \in [0, 2\pi]} \| x - e^{i\phi} x_0 \|,
\]

It is shown in [CLS14] that for \( m \geq c_0 n \log(n) \), the following holds with high probability:

\[
\text{dist}(z_0, x_0) \leq \frac{1}{8} \| x_0 \|.
\]  

(9.15)

115
Moreover for $\mu_\tau = \mu$, for all $\tau = 1, 2, \ldots$, and assume $\mu \leq c_1/n$, starting from that initial solution $z_0$, we have with high probability:
\[
\text{dist}(z_\tau, x_0) \leq \left(1 - \frac{\mu}{4}\right)^{\tau/2} \text{dist}(z_0, x_0) \leq \frac{1}{8} \left(1 - \frac{\mu}{4}\right)^{\tau/2} ||x_0||.
\] (9.16)

Setting $\mu = \frac{\epsilon}{n}$, we obtain an $\epsilon$-recovery, namely, $\text{dist}(z, x_0) \leq \epsilon ||x_0||$, in $O(n \log(\frac{1}{\epsilon}))$ iterations.

### 9.3 Summary of Our Main Results and Techniques

In this chapter, we propose and study a quantization scheme and a corresponding recovery procedure. In particular we investigate the properties of our approach towards: 1) the phase recovery problem from severely perturbed measurements, and 2) the improvement of the AM approach discussed in the previous section.

#### 9.3.1 Phase Recovery from Severely Perturbed Measurements

We investigate the phase recovery problem in the case in which we have at disposal measurements of the form
\[
b_i = \theta(|\langle a_i, x_0 \rangle|^2), \quad i = 1, \ldots, 2m,
\] (9.17)
where $a_i$ are sensing vectors as in (9.1) and $\theta$ is a possibly unknown rank preserving transformation. In particular we are interested to situations where $\theta$ models a distortion, e.g. $\theta(s) = \tanh(\alpha s)$, $\alpha \in \mathbb{R}_+$, or an additive noise $\theta(s) = s + \nu$, where $\nu$ is a stochastic noise, such as an exponential or Poisson noise. As we noted before the recovery problem from severely perturbed intensity values seems hopeless, and indeed the key in our approach is a quantization scheme based on comparing pairs of phase-less measurements. More precisely for each pair $b^1, b^2$ of measurements of the form (9.17) we define $y \in \{-1, 1\}$ as $y = \text{sign}(b^1 - b^2)$. This one-bit quantization scheme draws inspiration from ideas in one-bit compressive sensing [BB08, PV13b, PV13a, JLBB13], but the fundamental difference is that our approach crucially depends on the comparison of two measurements: one-bit measurements involve the spacing between the order statistics of exponentially distributed random variables $b^1$ and $b^2$. Indeed, this will be a key fact in our analysis. While phase-recovery from one-bit phase-less measurements is in general a hard problem (see Section 9.4), we propose to consider a relaxation which reduces to a maximum eigenvalue problem induced by the matrix
\[
\hat{C}_m = \frac{1}{m} \sum_{i=1}^{m} y_i(a_i^1 a_i^{1*} - a_i^2 a_i^{2*}).
\] (9.18)

When compared to (9.9) we see that the $m$ phase-less measurements $b_i$ are replaced by their quantized counterpart $y_i$ (obtained from $2m$ phase-less measurements), and the term given by sensing vectors is now given by pairs of sensing vectors. Indeed, we prove in Section 9.5
that the expectation of $\hat{C}_m$ satisfies
\[ \mathbb{E}\hat{C}_m = \lambda x_0 x_0^*, \]
where $\lambda$ is a suitable constant which depends on $\theta$ and plays the role of a signal-to-noise ratio. Indeed, by studying the concentration properties of the matrix $\hat{C}_m$, we show that, for a given accuracy $\epsilon \in [0, 1]$, if $O(\frac{n}{\epsilon^2 \lambda})$ pairs of measurements are available, then the solution of the above maximum eigenvalue problem satisfies
\[ \|\hat{x}_m - x_0 e^{i\phi}\|^2 \leq \epsilon, \]
where $\phi \in [0, 2\pi]$ is a global phase. It is worth noting here that the signal can be recovered up to a scaling factor from one bit measurements, but this is not a problem since our goal is to recover the missing phase.

### 9.3.2 One-Bit Phase Retrieval and Alternating Minimization

A key difference between the matrix $\hat{C}_m$ of SubExpPhase in Eq. (9.9) and the one of 1bitPhase in Eq. (9.18) is that, the one-bit measurements $y_i$ are bounded by one, and the unquantized measurements $b_i$ are bounded by $\log(n)$. Quantization implies boundedness of the measurements, which does not depend on the dimension $n$, this leads to improved concentration results of the Hermitian matrix $\hat{C}_m$ around its expectation.

Another advantage of our quantization scheme, goes to the pairing, that allows for an unbiased estimation of $x_0$. Indeed $\mathbb{E}(\hat{C}_m)$ is a rank one matrix in the case of 1bitPhase as shown in equation (9.29), as opposed to $\mathbb{E}(\hat{C}_m)$ in the case of SubExpPhase which is full rank as shown in equation (9.11). While SubExpPhase allows a biased estimation $x_0$, 1bitPhase allows for an unbiased estimation of $x_0$. Having a rank one matrix in expectation allows us to use matrix concentration inequalities taking into account the intrinsic dimension of the matrix [Tro12], the latter allows us to get improved sample complexity bounds. This motivates considering the effect of using the solution of the one-bit phase retrieval to initialize the alternating minimization procedure considered in [NJS13]. Indeed, leveraging results from [NJS13], we show in Section 9.6 that, provided with the one-bit retrieval initialization, the alternating minimization algorithm globally converges (with high probability) to the true vector $x_0$, and if
\[ m \geq cn \left( 1 + \log \frac{1}{\epsilon} \log \log \frac{1}{\epsilon} \right), \quad (9.19) \]
then $\|x_{t_0} - e^{i\phi}x_0\|_2 \leq \epsilon$. Comparing to (9.10), we see that the sample complexity depends now is linear in the dimension $n$. Similar improvements hold for the wirtinger flow approach of [CLS14]. Quantization can be seen as playing the role of a preconditioning that enhances the sample complexity of the alternating minimization.
Table 9.1: Comparison of the sample and computational complexity of different phase-retrieval schemes.

### 9.4 One-Bit Phase Retrieval

In this section, we set up the one-bit approach to phase-retrieval and state our main results on robust phase recovery.

#### 9.4.1 Quantization and Recovery

Unlike in compressive sensing, in our context the intensities are non-negative, and thus we cannot rely on only one measurement to build a quantizer.

**Definition 18** (One-bit quantizer). Let $A = (a^1, a^2)$, where $a^1, a^2$ are i.i.d. complex Gaussian vectors $\mathcal{N}(0, \frac{1}{2}I_n) + i\mathcal{N}(0, \frac{1}{2}I_n)$. For $x_0 \in \mathbb{C}^n$, a one bit quantizer is given by

$$Q_A : \mathbb{C}^n \to \{\pm 1\}, \quad Q_A(x_0) = \text{sign} \left( |\langle a^1, x_0 \rangle|^2 - |\langle a^2, x_0 \rangle|^2 \right).$$

**Definition 19** (Quantized phase-less measurements). Let $\{A_i = (a_i^1, a_i^2)\}_{1 \leq i \leq m}$ be $2m$ i.i.d. gaussian complex vectors in $\mathbb{C}^n$, and $Q_A(x_0)$ as in Def 22. The Quantized Phase-less sensing is given by $Q : \mathbb{C}^n \to \{\pm 1\}^m$, $Q(x_0) = (Q_{A_1}(x_0), \ldots, Q_{A_m}(x_0))$.

In this chapter, we are interested in recovering $x_0$ from its quantized phase-less measurements $y = (y_1, \ldots y_m) = Q(x_0) = (Q_{A_1}(x_0), \ldots, Q_{A_m}(x_0))$. It is easy to see that the recovery problem has the form,

$$\text{find } x, \quad \text{subject to } y_i \left( |\langle a_i^1, x \rangle|^2 - |\langle a_i^2, x \rangle|^2 \right) \geq 0, \quad i = 1, \ldots m, \quad ||x||_2 = 1. \quad (9.20)$$

Indeed, as in one-bit compressive sensing, we cannot hope to recover the norm of the vector from inequality constraints, hence the norm one constraint. Problem (9.20) can be equivalently written as the following quadratically constrained problem,

$$\text{find } x, \quad \text{subject to } x^* y_i (a_i^1 a_i^{1*} - a_i^2 a_i^{2*}) x \geq 0, \quad i = 1, \ldots m, \quad ||x||_2 = 1. \quad (9.21)$$

The above problem is a non-convex Quadratically Constrained Quadratic Program (QCQP) and can be shown to be NP-hard in general [DB03]. We propose to consider the following
relaxation,
\[
\max_x x^* \left( \frac{1}{m} \sum_{i=1}^{m} y_i (a_i^1 a_i^1 - a_i^2 a_i^2) \right) x, \quad \text{subject to } \|x\|_2^2 = 1. \tag{9.22}
\]

The 1BitPhase problem is obtained noting that the above problem can be rewritten as the the maximum eigenvalue problem,
\[
\max_{x \text{ s.t. } \|x\|_2=1} x^* \hat{C}_m x, \tag{9.23}
\]
defined by the matrix
\[
\hat{C}_m = \frac{1}{m} \sum_{i=1}^{m} y_i (a_i^1 a_i^1 - a_i^2 a_i^2).
\]

As we comment in the following remark the 1BitPhase approach is inspired by one-bit compressed sensing.

**Remark 6** (Quantization in Compressive Sensing: One bit CS). Non-linear, quantized measurements have been recently considered in the context of one-bit compressive sensing\(^1\). Here, binary (one-bit) measurements are obtained applying, for example, the “sign” function to linear measurements. More precisely, given \(x_0 \in \mathbb{R}^n\), a measurement vector is given by \(y = (y_1, \ldots, y_m)\), where \(y_i = \text{sign}(\langle a_i, x \rangle)\) with \(a_i \sim \mathcal{N}(0, I_n)\) independent Gaussian random vectors, for \(i = 1, \ldots, m\). It is possible to prove, see e.g. [PV13b], that, for a signal \(x_0 \in K \cap \mathbb{B}^n\) (\(\mathbb{B}^n\) is the unit ball in \(\mathbb{R}^n\)), the solution \(\hat{x}_m\) to the problem
\[
\max_{x \in K \cap \mathbb{B}^n} \frac{1}{m} \sum_{i=1}^{m} y_i \langle a_i, x \rangle, \tag{9.24}
\]
satisfies \(\|\hat{x}_m - x_0\|^2 \leq \frac{\delta}{\sqrt{\frac{n}{2}}}\), \(\delta > 0\), with high probability, as long as \(m \geq C \delta^{-2} \omega(K)^2\) [PV13b]. Here \(\omega(K) = \mathbb{E} \sup_{x \in K-K} \langle w, x \rangle\) denotes the Gaussian mean width of \(K\). The 1BitPhase approach shows that a relaxation similar to problem 9.24 allows to perform phase recovery for a suitably defined quantization of phase-less linear measurements.

Before studying the recovery guarantees for the solution of 1BitPhase we briefly discuss a geometric intuition underlying the method.

### 9.4.2 Geometric Intuition

To understand the geometric intuition of one bit phase retrieval, we first consider the feasibility problem (9.20). Each measurements pair \((a_i^1, a_i^2)\) defines a hyperbolic paraboloid \(z = x^*(a_i^1 a_i^1 - a_i^2 a_i^2) x\). The feasible zone is defined by the constraints \(y_i (x^*(a_i^1 a_i^1 - a_i^2 a_i^2) x) \geq 0\),

\(^1\)See [http://dsp.rice.edu/1bitCS/](http://dsp.rice.edu/1bitCS/) for an exhaustive list of references.
which enforce the geometric consistency with the bit $y_i$. In other words each constraint says that $x_0$ is in the region of the space where the sign of the corresponding hyperbolic paraboloids is $y_i$. Note that each hyperbolic paraboloid is symmetric with respect to the origin, thus the feasible region is also symmetric with respect to the origin. When we add more constraints we have that $x_0$ lies in the intersection of such symmetric feasible regions and the unit sphere. This intersection is also symmetric, thus we can solve the phase retrieval up to global sign flip (in the real valued case). As we mentioned before, the feasibility problem is a non convex QCQP which is NP hard in general and our relaxation (9.23) can be seen as requiring the geometric consistency with one bit measurements on average, rather than individually as in the feasibility problem. In figure 9-1 we plot the level sets of the objective function of the 1bitPhase Problem (9.23) for $x_0 = (1, 0)$. We see that the objective function achieves its maximum values (in red) in a symmetric region. This region intersects with the sphere in two regions close to the points $x_0$ and $-x_0$. Thus we are able to recover the phase up to global sign and scaling from single bit measurements.

9.4.3 One bit Phase Retrieval from Distorted and Noisy measurements

In the following we assume that the intensities are undergoing an unknown non linearity $\theta$, that is we observe,

$$(b_i^1, b_i^2) = (\theta(|\langle a_i^1, x_0 \rangle|^2), \theta(|\langle a_i^2, x_0 \rangle|^2)), \quad i = 1, \ldots, m.$$
Thus the quantized measurements are,

\[ y_i = Q_{\Lambda_i}^\theta(x_0) = \text{sign} \left( \theta(| \langle a_i^1, x_0 \rangle |^2) - \theta(| \langle a_i^2, x_0 \rangle |^2) \right), \quad i = 1 \ldots m. \]  

(9.25)

For instance clipping can be modeled by a sigmoid,

\[ \theta(z) = \tanh(\alpha z), \quad z > 0. \]

where the parameter \( \alpha \) controls how severe is the distortion. An additive noise (before quantization) can be modeled by,

\[ \theta(z) = z + \nu, \]

(9.26)

where \( \nu \sim \text{Exp}(\gamma) \) is a stochastic exponential noise with mean \( \mu = \frac{1}{\gamma} \) and variance \( \sigma = \frac{1}{\gamma^2} \).

When the intensities are contaminated with poisson noise we have,

\[ \theta(| \langle a, x_0 \rangle |^2) = \mathcal{P}_\eta (| \langle a, x_0 \rangle |^2), \quad (9.27) \]

where \( \mathcal{P}_\eta \) is a poisson noise, such that :

For \( z, \eta > 0, \quad \theta(z) = \mathcal{P}_\eta(z) = p, \quad \text{where } p \sim \text{Poisson} \left( \frac{z}{\eta} \right). \)

We shall make one assumption on the non linearity \( \theta \),

\[ \lambda = \mathbb{E}(\text{sign}(\theta(E_1) - \theta(E_2))(E_1 - E_2)) > 0, \]

(9.28)

where \( E_1, E_2 \) are two independently distributed exponential random variables. To see why this assumption is natural, notice that \( | \langle a, x_0 \rangle |^2 \sim \text{Exp}(1) \) if \( a \sim \mathcal{CN}(0, I_n) \) and \( ||x_0|| = 1 \), thus

\[ \mathbb{E}(y_i(| \langle a_i^1, x_0 \rangle |^2 - | \langle a_i^2, x_0 \rangle |^2)) = \mathbb{E}(\text{sign}(\theta(E_1) - \theta(E_2))(E_1 - E_2)) = \lambda > 0. \]

Then the above assumption simply means that the one bit measurements preserve robustly the ranking of the intensities. For example, such an assumption is trivially satisfied whenever \( \theta \) is increasing. In case \( \theta(z) = z \), \( \lambda \) achieves its maximal value,

\[ \lambda = \mathbb{E}(\text{sign}(E_1 - E_2)(E_1 - E_2)) = \mathbb{E}|E_1 - E_2| = 1, \]

since \( |E_1 - E_2| \sim \text{Exp}(1) \). For different models of observation, the value of \( \lambda \) is given in Lemma 10, which shows that that \( \lambda \) plays the role of a Signal to Noise Ratio (SNR).
Figure 9-2: \( \lambda \) as a signal to noise ratio: \( \lambda \) versus different parameters of various observation model \( \theta \). \( \lambda \) decreases as noise and distortion levels increase.

We see in Figure 9-2 that \( \lambda \) indeed decreases as the level of noise and distortion increases. The following Lemma shows how the value of \( \lambda \) depends on the observation model \( \theta \), and how \( \lambda \) relates to noise and distortion levels.

**Lemma 10.** The values of \( \lambda \) for different observation models \( \theta \) are given in the following:

1. **Noiseless setup:** \( \theta(z) = z \), \( \lambda = 1 \).

2. **Exponential Noise:** \( \theta(z) = z + \nu, \nu \) is an exponential random variable with variance \( \sigma \), \( \lambda = \frac{1+2\sqrt{\sigma}}{1+\sqrt{\sigma}} \).

3. **Poisson Noise:** \( \theta(z) = \mathcal{P}_\eta(z), \mathcal{P}_\eta(z) = p \) \( p \sim \text{Poisson} \left( \frac{z}{\eta} \right) \), \( \lambda = \mathbb{E}(\text{sign}(S(E_1, E_2))(E_1 - E_2)) \), \( S(E_1, E_2) \sim \text{Skellam} \left( \frac{E_1}{\eta}, \frac{E_2}{\eta} \right), E_1, E_2 \sim \text{Exp}(1) \). \( \lambda \) is a decreasing function in \( \eta \).

4. **Distortion setup:** \( \theta(z) = \tanh(\alpha z) \), \( \lambda = \mathbb{E}(|E_1 - E_2| \text{sign}(1-\tanh(\alpha E_1) \tanh(\alpha E_2))) \), \( E_1, E_2 \sim \text{Exp}(1) \). \( \lambda \) is a decreasing function in \( \alpha \).

From Lemma 10, we see that \( \lambda \) achieves its maximum value 1, in the noiseless case. \( \lambda \) interestingly captures the SNR as it decreases with noise and distortion levels.

### 9.4.4 Main Results for One Bit Phase Retrieval

The following theorem describes the recovery guarantees for the solution \( \hat{x}_m \) of problem 1bitPhase (9.23).

**Theorem 8** (One bit Recovery). For \( x_0 \in \mathbb{C}^n, \|x_0\| = 1 \). Assume \( y_1 \ldots y_m \), follows the model given in (9.25). Then for any \( \epsilon \in ]0,1[, \) we have with a probability at least \( 1 - O(me^{-2n}) \),

\[
\text{for } m \geq \frac{C}{\epsilon^2 \lambda n}, \quad \|\hat{x}_m \hat{x}_m^* - x_0 x_0^*\|_F^2 \leq \epsilon,
\]
where $\lambda$ is given in (9.28). Let $\phi \in [0, 2\pi]$, we have therefore with a probability at least $1 - O(me^{-2n})$,

$$
\| \hat{x}_m - x_0 e^{i\phi} \|^2 \leq \epsilon.
$$

For the simple model model where $\theta(z) = z$, $\lambda = 1$. Theorem 8 implies that if $m = O(n)$ (so that the total number of measurements is $2m$), then $\hat{x}_m$ is an $\epsilon$-estimate of $x_0$, up to a global phase $\phi$. In Corollary 3 we specify the above theorem to the noisy model (9.26).

**Corollary 1** (One bit Recovery/ Noise). For $x_0 \in \mathbb{C}^n$, $||x_0|| = 1$, and $\epsilon > 0$. Assume $y_1 \ldots y_m$, follows the noisy model given in (9.25), for $\theta(z) = z + \nu$, $\nu \sim \text{Exp}(\gamma)$. Where $\nu$ is an exponential noise with variance $\sigma = \frac{1}{\gamma^2}$. Then for any $\epsilon \in [0, 1]$, we have with a probability at least $1 - O(me^{-2n})$,

$$
\| \hat{x}_m \hat{x}_m^* - x_0 x_0^* \|_F^2 \leq \epsilon.
$$

In other words, under an exponential noise we have:

$$
\| \hat{x}_m \hat{x}_m^* - x_0 x_0^* \|_F^2 \leq C \sqrt{\frac{n (1 + \sqrt{\sigma})^2}{m (1 + 2\sqrt{\sigma})}}.
$$

A similar result holds for Poisson noise for a different value of $\lambda$ given in Lemma 10.

Beyond robustness to noise, another desirable feature for phase retrieval from phase-less measurements, is the robustness to distortions of the values of intensities. Is it possible to retrieve the phase from intensities values that are undergoing clipping for instance?

**Corollary 2** (One bit Recovery/ Distortion). For $x_0 \in \mathbb{C}^n$, $||x_0|| = 1$, and $\epsilon > 0$. Assume $y_1 \ldots y_m$, follows the noisy model given in (9.25), for $\theta(z) = \tanh(\alpha z)$, $\alpha > 0$. Then for any $\epsilon \in [0, 1]$, we have with a probability at least $1 - O(me^{-2n})$,

$$
\| \hat{x}_m \hat{x}_m^* - x_0 x_0^* \|_F^2 \leq \epsilon,
$$

where $\lambda(\alpha) = \mathbb{E}(|E_1 - E_2| \text{sign}(1 - \tanh(\alpha E_1) \tanh(\alpha E_2)))$ is a decreasing function in $\alpha$.

The proof of the above results follow from a simple combination of Propositions 4, and 8 given in Section 9.5.

**9.5 Theoretical Analysis**

In this section we give the main steps of the proof of Theorem 8 for one bit phase Retrieval.
9.5.1 One Bit Phase Retrieval: Correctness and Concentration

In this section we state Propositions 4, and 8 which form the core of our analysis for one bit Phase Retrieval. The proofs are given in Section 9.9. We need the following preliminary definition.

Definition 20 (Risk and Empirical risk). Let \( x_0 \in \mathbb{C}^n, ||x_0|| = 1 \). For \( x \in \mathbb{C}^n \) such that \( ||x|| = 1 \), and \( A = \{a_1, a_2\} \) i.i.d. complex Gaussians, let

\[
\mathcal{E}^{x_0}(x) = x^* C x,
\]

where \( C = \mathbb{E}(y(a_1^*a_1 - a_2^*a_2)) \) and \( y = \text{sign}(\theta(|\langle a_1, x_0 \rangle|^2) - \theta(|\langle a_2, x_0 \rangle|^2)) \). Moreover, let

\[
\hat{\mathcal{E}}^{x_0}(x) = x^* \hat{C}_m x,
\]

where \( \hat{C}_m = \frac{1}{m} \sum_{i=1}^{m} y_i (a_1^*a_1 - a_2^*a_2), y_i = Q^\theta_{A_i}(x_0) \) and \( A_i = \{(a_1^i, a_2^i)\}, i = 1 \ldots m \) are i.i.d. complex Gaussians.

We first state Proposition 4, that provides a theoretical justification to the relaxation introduced in the formulation 1bitPhase in (9.23).

Proposition 4 (Correctness in Expectation). Let \( x_0 \in \mathbb{C}^n, \) such that \( ||x_0|| = 1 \). The following statements hold:

1. Let \( y = Q^\theta_A(x_0) \), as given in equation (9.25), then \( C \) is a rank one matrix, and moreover we have,

\[
C = \mathbb{E}(y(a_1^*a_1 - a_2^*a_2)) = \lambda x_0 x_0^*,
\]

where \( \lambda = \mathbb{E}(\text{sign}(\theta(E_1) - \theta(E_2)))(E_1 - E_2), \) and \( E_1, E_2 \) iid \( \sim \text{Exp}(1) \).

2. For all \( x \in \mathbb{C}^n, ||x|| = 1 \), we have the following equality,

\[
\mathcal{E}^{x_0}(x) = x^* C x = \lambda |\langle x_0, x \rangle|^2.
\]

3. \( x_0 \) is an eigen vector of \( C \) with eigen value \( \lambda \),

\[
Cx_0 = \lambda x_0.
\]

4. The maximum eigenvector of \( C \) is of the form \( x_0 e^{i\phi} \), where \( \phi \in [0, 2\pi] \). The maximum eigenvalue is given by \( \lambda \).

Proposition 4 suggests that \( x_0 \) can be recovered as the maximum eigenvector of the matrix \( C \). Moreover \( C \) is a rank one matrix. The corresponding eigenvalue \( \lambda \), quantifies the quality of the recovery with respect to the observation model. Lemma 10 shows how the value of \( \lambda \) depends on the observation model \( \theta \), and how \( \lambda \) relates to noise and distortion.
levels.

When compared to Equation (9.11), we see that \( C \) is a full rank matrix in the case of \( \text{SubExpPhase} \), while it is a rank one matrix in the case \( \text{1bitPhase} \) as stated in Equation (9.29). Hence \( \text{1bitPhase} \) allows an unbiased estimation of the point of interest.

The empirical problem amounts therefore to finding \( \hat{x}_m \) the maximum eigenvector of the matrix

\[
\hat{C}_m = \frac{1}{m} \sum_{i=1}^{m} y_i (a_i^1 a_i^{1,*} - a_i^2 a_i^{2,*}).
\]

Lemma 11 (Comparison Inequality). Let \( \hat{x}_m \) be the maximum eigenvector of \( \hat{C}_m \), we have:

\[
\frac{\lambda}{2} \left\| \hat{x}_m \hat{x}_m^* - x_0 x_0^* \right\|_F^2 \leq 2 \left\| \hat{C}_m - C \right\|.
\]

Lemma 11 shows that the concentration of the self adjoint matrix \( \hat{C}_m \) around \( C \) controls the sample complexity of the recovery. Interestingly in One Bit Phase retrieval the matrix \( C \) is a rank one matrix. Recent results in matrix concentration inequalities allow us to get improved concentration results taking into account the intrinsic dimension. Before studying the concentration of \( \hat{C}_m \) around its mean we pause to give a precise definition of the intrinsic dimension and introduce the matrix Bernstein concentration inequality with intrinsic dimension.

Definition 21 (Intrinsic Dimension [Tro12]). For a positive-semidefinite matrix \( A \), the intrinsic dimension is defined as:

\[
\text{intdim}(A) = \frac{\text{tr} A}{||A||}.
\]

Theorem 9 ([Tro12]). Consider a finite sequence \( X_i \) of independent centered self adjoint random \( n \times n \) matrices. Assume we have for some number \( R \) and a positive semidefinite matrix \( V \) that:

\[
||X_i|| \leq R \text{ almost surely } \quad \mathbb{E} \sum_i X_i^2 \preceq V.
\]

Then, for every \( t > \sqrt{n} \frac{R}{3} \), we have:

\[
\mathbb{P} \left( \left\| \sum_i X_i \right\| > t \right) \leq 4 \text{intdim}(V) \exp \left( \frac{-t^2/2}{||V|| + Rt/3} \right).
\]

When compared with Bernstein inequality (Theorem 10) \( \text{intdim}(V) \) replaces the dimension \( n \). Note that \( 1 \leq \text{intdim}(V) \leq \text{rank}(V) \leq \text{dim}(V) = n \), hence this result improves the sample complexity bound.

We are now ready to state the concentration result using the fact that \( C \) is a rank one matrix.
Proposition 5 (Concentration). For sufficiently large $n$ we have that,

$$\text{for } m \geq \frac{cn}{\lambda^2}, \quad \left\| \hat{C}_m - C \right\| \leq \epsilon \lambda \text{ with probability at least } 1 - O(me^{-2n}).$$

9.6 One Bit Phase Retrieval as an Initialization to non convex approaches

Note that $b_i, i = 1 \ldots m$ are exponential random variable thus we call that initialization Sub-exponential Initialization. The concentration of $\hat{C}_m$ around $C$, depends upon the boundedness of $b_i$ and $a_i$ by the non commutative matrix Bernstein inequality (Theorem 10). We have with high probability that

$$b_i ||a_i||^2 \leq 4 \log(m)n,$$

thus we have a sample complexity of $O(n \log^3(n))$ due to the extra contribution of $b_i$ with a $\log(n)$ term. Recall that the solution of one bit phase retrieval and weighted one bit phase retrieval is the maximum eigen vector of

$$\hat{C}_m = \frac{1}{m} \sum_{i=1}^{m} y_i (a_i^1 a_i^{1*} - a_i^2 a_i^{2*}).$$

The measurements $y_i$ are bounded by 1 and do not affect the bound. As discussed earlier, while SubExpPhase allows a biased estimation $x_0$, 1bitPhase allows for an unbiased estimation of $x_0$, thanks to the pairing in the quantized measurement scheme that removes the bias term. Thus, the sample complexity reduces to only $O(n)$ pairs of measurements for phase retrieval via one-bit measurements. Thus we can initialize the alternating minimization with the solution of One Bit Phase and get a better sample complexity especially in high dimensions.

Geometric intuition. We see in Figure 9-3 that the levels sets of the objective of SubExpPhase initialization consists of a paraboloid, that is symmetric, hence it intersects the unity sphere in a symmetric zone thus phase retrieval is possible up to a global phase. Compared to one bit initialization, the levels sets are hyperbolic paraboloids (See Figure 9-1). Hyperbolic paraboloid are more "pointy" than paraboloid thus the surface of intersection with the sphere is smaller. The above discussion gives an intuition of the reasons behind the better sample complexity of one-bit phase retrieval.
Leveraging results from [NJS13] that does not depend on the initialization step, Lemma 12 shows that the greedy refinements of one bit solution, ensures convergence to the optimum with high probability and lower sample complexity than the ones obtained in [NJS13]. Let 
\[ \text{dist}(x, x_0) = \sqrt{1 - \left( \frac{||x, x_0||}{||x|| ||x_0||} \right)^2} \], the following steps give a glimpse on the reasons behind the success of the approach outlined in Algorithm 3, when initialized with the one bit solution:

1. Let \( \hat{x}_m \) be the maximum eigenvector of \( \hat{C}_m \) of 1bitPhase, then for \( 0 < \epsilon < 1 \), the following hold with high probability:

   For \( m \geq \frac{c}{\epsilon^2} n \), \( \text{dist}^2(\hat{x}_m, x_0) \leq \epsilon. \)

2. For \( t = 1 \ldots t_0 \), let \( \hat{x}_t \) be the sequence defined in Algorithm 3, where \( \hat{x}_0 \) is the maximum eigenvector of the matrix formed using samples \((b_0, A_0)\). The following holds with high probability [NJS13]:

   \[ \text{dist}(\hat{x}_t, x_0) \leq \frac{3}{4} \text{dist}(\hat{x}_{t-1}, x_0) \leq \left( \frac{3}{4} \right)^t \text{dist}(\hat{x}_0, x_0). \]

**Lemma 12** (Greedy Refinements). Let \( \hat{x}_m \) be the solution of One bit Phase Retrieval, and consider Algorithm 3 initialized with \( \hat{x}_m \) for all \( \epsilon, 0 < \epsilon < 1 \). Define \( x_{t_0} \) the output of Algorithm 3. For \( m = O \left( 2n(1 + \log \frac{1}{\epsilon} \log \log \frac{1}{\epsilon}) \right) \), we have \( ||x_{t_0} - x_0e^{i\phi}||_2 \leq \epsilon \) with high probability.
Similar results can be obtained for the Wirtinger flow approach as we change the initialization step to the one bit solution.

9.7 Computational Aspects

9.7.1 One bit Phase Retrieval Algorithms

A straightforward computation of the maximum eigenvector of the matrix $\hat{C}_m$ is expensive. One needs $O(n^2 m)$ operations to compute the matrix $\hat{C}_m$, that is $O(n^3)$, and then the computation of the first eigenvector requires $O(n^2)$ operations. The total computational cost is therefore $O(n^3 + n^2)$, and is dominated by the cost of computing $\hat{C}_m$.

An elegant method to avoid that overhead is the power method. The power method allows for the computation of the maximum eigenvector without having to compute the matrix $\hat{C}_m$, this reduce drastically the computational cost to $O(nm)$ at each iteration of the power method, that is $O(n^2)$. In the following we discuss an algorithm for one bit phase retrieval based one the power method:

1bitPhasePower: One bit Phase retrieval via the Power Method given in Algorithm 4.

Algorithm 4 1bitPhasePower

1: procedure 1BITPHASEPOWER($A$, $y$, $\epsilon$)
2: Initialize $r_0$ at random, $j = 1$.
3: while $||r_j - r_{j-1}|| > \epsilon$ or $j = 1$ do
4:   $r_j \leftarrow \frac{1}{m} \sum_{i=1}^{m} y_i (\langle a_i^1, r_{j-1} \rangle a_i^1 - \langle a_i^2, r_{j-1} \rangle a_i^2)$
5:   $\hat{\lambda} \leftarrow ||r_j||$
6:   $r_j \leftarrow \frac{r_j}{\hat{\lambda}}$
7:   $j \leftarrow j + 1$
8: end while
9: return $(\hat{\lambda}, r)$ \quad $\triangleright$ $(\hat{\lambda}, r)$ is an estimate of $(\lambda, x_0)$.
10: end procedure

9.8 Numerical Experiments

We consider a signal $x_0 \in \mathbb{C}^n$ which is a random complex Gaussian vector with i.i.d. entries of the form $x_0[j] = X + iY$, where $X, Y \sim \mathcal{N}(0, \frac{1}{2})$, $j = 1 \ldots n$. Let $n = 128$, $\epsilon = 0.25$. We consider $r = 1/(\epsilon^2)[\log(n)] = 64$ and set $m = rn$. So that the total number of measurements
Figure 9-4: Robustness of One Bit Phase Retrieval to extreme distortion: the error of recovery $1 - |\langle x, x_0 \rangle|^2$ versus the distortion level $\alpha$.

is $2m$. We assume that we measure distorted (clipped) measurements according to the model:

$$b^1_i = \tanh(\alpha|\langle a^1_i, x_0 \rangle|^2), \quad b^2_i = \tanh(\alpha|\langle a^2_i, x_0 \rangle|^2) \quad a^1_i, a^2_i \sim \mathcal{CN}(0, I_n), i = 1 \ldots m.$$ 

$\alpha$ corresponds to the level of distortion. The distortion is more severe as $\alpha$ increases. In figure 3 we plot the error of recovery $1 - \text{dist}(\hat{x}_m, x_0) := 1 - |\langle \hat{x}_m, x_0 \rangle|^2$. Where $\hat{x}_m$ is either the solution of 1bitPhase or SubExpPhase. We see that one bit phase retrieval robustly recovers the signal while traditional approaches (SubExpPhase for instance) fail under severe distortions.
9.9 Proofs of Lemmas 10 and 11, Propositions 4, and 8.

We give here the proof of Lemma 10:

Proof of Lemma 10. i. Noiseless:
\[ \lambda = \mathbb{E}(\text{sign}(E_1 - E_2)(E_1 - E_2)) = \mathbb{E}(|E_1 - E_2|) = 1, \text{ since } E_1 - E_2 \sim \text{Exp}(1). \]

ii. Noisy:
Exponential Noise:
Let \( y = \text{sign}((E_1 + \nu_1) - (E_2 + \nu_2)) \). Let \( L = E_1 - E_2 \), \( L \) follows a Laplace distribution with mean 0 and scale parameter 1:

\[ L \sim \text{Laplace}(0, 1). \]

Let \( N = \nu_1 - \nu_2 \), \( N \) follows a Laplace distribution, \( N \sim \text{Laplace}(0, \frac{1}{\gamma}) \). It follows that:

\[
\begin{align*}
\lambda &= \mathbb{E}_{L,N}(\text{sign}(L + N)L) \\
&= \mathbb{E}_L((1 - 2\mathbb{P}_N(N \leq -L))L) \\
&= \mathbb{E}_L((1 - 2\mathbb{F}_N(-L))L) \\
&= \mathbb{E}_L\left\{ \left(1 - 2\left(\frac{1}{2} + \frac{1}{2}\text{sign}(-L)(1 - \exp(-\gamma|L|))\right)\right) L \right\} \\
&= \mathbb{E}_L(\text{sign}(L)(1 - \exp(-\gamma|L|))L) \\
&= \mathbb{E}_L[L|1 - \exp(-\gamma|L|)] \\
&= 1 - \int_0^{\infty} z\exp(-\gamma z)\exp(-z)dz \\
&= 1 - \frac{1}{(1 + \gamma)^2} > 0.
\end{align*}
\]

Let \( \sigma = \frac{1}{\gamma^2} \) be the variance of the exponential noise. We conclude that:

\[ \lambda = \frac{1 + 2\sqrt{\sigma}}{(1 + \sqrt{\sigma})^2}. \]

Poisson Noise:

\[ \lambda = \mathbb{E}(\text{sign}(p_1 - p_2)(E_1 - E_2)), \quad p_1|E_1 \sim \text{Poisson}\left(\frac{E_1}{\eta}\right), \quad p_2|E_2 \sim \text{Poisson}\left(\frac{E_2}{\eta}\right). \]

We know that:

\[ S(E_1, E_2) = p_1 - p_2|E_1, E_2 \sim \text{Skellam}\left(\frac{E_1}{\eta}, \frac{E_2}{\eta}\right). \]

Hence:

\[ \lambda = \mathbb{E}(\text{sign}(S(E_1, E_2))(E_1 - E_2)). \]
iii. Distortion:

\[ y = \text{sign}(\tanh(\alpha E_1) - \tanh(\alpha E_2)) \]
\[ = \text{sign}(\tanh(\alpha (E_1 - E_2))) (1 - \tanh(\alpha E_1) \tanh(\alpha E_2))) \]
\[ = \text{sign}(\tanh(\alpha (E_1 - E_2))) \text{sign} (1 - \tanh(\alpha E_1) \tanh(\alpha E_2))) \]
\[ = \text{sign}(E_1 - E_2) \text{sign} (1 - \tanh(\alpha E_1) \tanh(\alpha E_2))) \]

\[ \lambda = \mathbb{E}(y(E_1 - E_2)) = \mathbb{E} (\text{sign} (1 - \tanh(\alpha E_1) \tanh(\alpha E_2)) | E_1 - E_2) > 0. \]

The proof of correctness in expectation (Proposition 4) of 1bitPhase is given in the following:

**Proof of Proposition 4.** 1) Let \( e_j, j = 1 \ldots n, \) be the canonical basis. Note \( a_j^l = \langle a_j, e_l \rangle, j = 1, 2, \) and \( \ell = 1 \ldots n. \) By rotation invariance of Gaussians we can consider \( x_0 = e_1 = (1, 0 \ldots 0). \) Hence \( y = \text{sign}(\theta(|\alpha_1|^2) - \theta(|\alpha_2|^2)). \) Let \( E_1 = |\alpha_1|^2, E_2 = |\alpha_2|^2, E_1 \) and \( E_2 \) are iid exponential random variables \( \text{Exp}(1). \) For \( z \in \mathbb{C}, \) we note \( \bar{z} \) the complex conjugate of \( z. \)

It follows that:

\[ C \mathrel{=} \mathbb{E}(y(a_1 a_1^* - a_2 a_2^*)) = \mathbb{E}\left( \sum_{k, \ell} y(a_k^1 a_\ell^1 - a_k^2 a_\ell^2) e_k e_\ell^* \right) = \mathbb{E} (\text{sign}(\theta(|\alpha_1|^2) - \theta(|\alpha_2|^2))(|\alpha_1|^2 - |\alpha_2|^2)) e_1 e_1^* = \lambda e_1 e_1^*, \]

where the last equalities follow since for \( k, \ell \neq 1 \) we have that \( y \) and the random variable \( a_k^1 a_\ell^1 - a_k^2 a_\ell^2 \) are independent, and moreover:

1. For \( k \neq \ell \) we have:
   \[ \mathbb{E} (y(a_k^1 a_\ell^1 - a_k^2 a_\ell^2)) = \mathbb{E}(y) \mathbb{E}(a_k^1 a_\ell^1 - a_k^2 a_\ell^2) = \mathbb{E}(y) \left( \mathbb{E}(a_k^1) \mathbb{E}(a_\ell^1) - \mathbb{E}(a_k^2) \mathbb{E}(a_\ell^2) \right) = 0. \]
   Since \( \mathbb{E}(a_j^l) = 0, j \in \{1, 2\} \) \( \forall \ell, \ell \neq 1. \)

2. For \( k = \ell \) we have:
   \[ \mathbb{E} (y(a_k^1 a_k^1 - a_k^2 a_k^2)) = \mathbb{E}(y) \mathbb{E}(|\alpha_k|^2 - |\alpha_k|^2) = 0, \] since \( \mathbb{E}(|\alpha_k|^2) = \mathbb{E}(|\alpha_k|^2) = 1, \) \( \forall k, k \neq 1. \)

(2), (3), and (4) follow immediately from (1). \( \square \)

The proof of the comparison inequality relating the concentration of the empirical matrix \( \hat{C}_m \) around its mean, to the distance to the optimum is given in the following:
Proof of Lemma 11. For $x \in \mathbb{C}^n, \|x\| = 1$, let $\mathcal{E}^{x_0}(x) = x^* C x$, and $\hat{\mathcal{E}}^{x_0}(x) = x^* \hat{C}_m x$.

$$
\mathcal{E}^{x_0}(x_0) - \mathcal{E}^{x_0}(x) = \lambda - \lambda |\langle x_0, x \rangle|^2 = \frac{\lambda}{2} \|xx^* - x_0x_0^*\|_F^2.
$$

Let $\tilde{x}_m = \arg \max_{x, \|x\| = 1} \hat{\mathcal{E}}^{x_0}(x)$, we have: $\mathcal{E}^{x_0}(x_0) - \mathcal{E}^{x_0}(\tilde{x}_m) = \mathcal{E}^{x_0}(x_0) - \hat{\mathcal{E}}^{x_0}(x_0) + \hat{\mathcal{E}}^{x_0}(\tilde{x}_m) - \mathcal{E}^{x_0}(\tilde{x}_m)$. Noticing that the term $\hat{\mathcal{E}}^{x_0}(x_0) - \hat{\mathcal{E}}^{x_0}(\tilde{x}_m)$ is non-positive in light of the definition of $\tilde{x}_m$, we have finally: $\mathcal{E}^{x_0}(x_0) - \mathcal{E}^{x_0}(\tilde{x}_m) \leq 2 \sup_{x, \|x\| = 1} \|\hat{\mathcal{E}}^{x_0}(x) - \mathcal{E}^{x_0}(x)\| = 2 \|\hat{C}_m - C\|$. \hfill \Box

In the following, using concentration inequalities, we assess the sample complexity of $1$bitPhase:

Proof of Proposition 8. Let $X_i = \frac{1}{m} (y_i (a_i^1 a_i^2*) - a_i^1 a_i^2* x_0^*) - \lambda x_0 x_0^*).$ We would like to get a bound on $|\sum_{i=1}^m X_i|$, the main technical issue is the fact that $|X_i|$ are not bounded almost surely. We will address this issue by rejecting samples outside the ball of radius $\sqrt{M}$, where $M$ is defined in the following.

Let $M = 2n(1 + \beta)^2$. Let $E = \{(a_1, a_2), \|a_1\|^2 \leq M \text{ and } \|a_2\|^2 \leq M\}$. Let

$$(a_{i_1}^1, a_{i_2}^1) = (a_{i_1}^1, a_{i_2}^2) \text{ if } (a_{i_1}^1, a_{i_2}^2) \in E \text{ and 0 otherwise.}$$

Let $\tilde{y}_i = \text{sign} (\theta(|\langle a_{i_1}^1, x_0 \rangle|^2) - \theta(|\langle a_{i_2}^2, x_0 \rangle|^2)) \text{ if } (a_{i_1}^1, a_{i_2}^2) \in E \text{ and 0 otherwise.}$ Let

$$
\tilde{C}_m = \frac{1}{m} \sum_{i=1}^m \tilde{y}_i (a_{i_1}^1 a_{i_1}^2* - a_{i_2}^2 a_{i_2}^2*) \quad \tilde{C} = \text{E}(\tilde{C}_m).
$$

Note that $\tilde{C}_m$ is the sum of bounded random variable, so that we can use the matrix Bernstein inequality given in Theorem 9, in order to bound $\|\tilde{C}_m - \tilde{C}\|$. On the other hand by the triangular inequality we have:

$$
\|C_m - C\| \leq \|C_m - \tilde{C}_m\| + \|\tilde{C}_m - \tilde{C}\| + \|\tilde{C} - C\| \quad (9.32)
$$

Bounding $\|C_m - \tilde{C}_m\|$:

Note that $\|a\|^2 \sim \chi^2_{2n}, \|a\|$ is a Lipchitz function of Gaussian with constant one. A Gaussian concentration bound implies,

$$
P(\|a_i\|^2 \geq (\sqrt{2n} + t)^2) \leq e^{-\frac{t^2}{2}}. \quad (9.33)
$$

Setting $t = \beta \sqrt{2n}$, it follows that: $P(\|a_i\|^2 \geq 2n(1 + \beta)^2) \leq e^{-\beta^2 n}$. Setting $t = \beta \sqrt{2n}$, it follows that: $P(\|a_i\|^2 \geq 2n(1 + \beta)^2) \leq e^{-\beta^2 n}$.

$$
P(\max_{i=1, \ldots, m, j=1, 2} \|a_{ij}\|^2 > M) \leq 2m P(\|a\|^2 > M) \leq 2m e^{-\beta^2 n}.
$$

132
It follows that:
\[
\left\| C_m - \hat{C}_m \right\| = 0 \text{ with probability at least } 1 - 2me^{-\beta^2n}.
\]

**Bounding \( \left\| \hat{C} - C \right\| :**

By the rotation invariance of Gaussian we can assume \( x_0 = (1, 0, \ldots, 0) \).
The off diagonal terms of \( \mathbb{E}((\hat{a}^1_a \hat{a}^1_1 - \hat{a}^2_a \hat{a}^2_1)^*) \) are zero. The same holds for \( \mathbb{E}(y(a^1_a a^1_1 - a^2_a a^2_1)^*) \).
The only term that is non zero on the diagonal is the first one.

\[
\left\| \hat{C} - C \right\| = \mathbb{E}(y(|a^1_1|^2 - |a^2_1|^2)^1_{(a^1_a, a^2_a) \notin E}) \\
\leq \left( \mathbb{E}(y^2(|a^1_1|^2 - |a^2_1|^2)^2) \right)^{\frac{1}{2}} \left( \mathbb{E}(1_{E^c}) \right)^{\frac{1}{2}} \\
= \left( \mathbb{E}(|a^1_1|^4 + |a^2_1|^4 - 2|a^1_1|^2|a^2_1|^2) \right)^{\frac{1}{2}} \sqrt{p(E^c)} \\
\leq \sqrt{2} + 2 - 2\sqrt{2}e^{-\beta^2n} \\
= 2e^{-\beta^2n/2}.
\]

**Bounding \( \left\| \hat{C}_m - \hat{C} \right\| :**

It is easy to see that: \( \hat{C} = \lambda e_1 e^*_1, \hat{\lambda} = \lambda - \mathbb{E} \left( \text{sign}(\theta(|a^1_1|^2) - \theta(|a^2_1|^2))(|a^1_1|^2 - |a^2_1|^2)1_{(a^1,a^2) \notin E} \right) \).
Let
\[
\hat{X}_i = \frac{1}{m} \left( \hat{y}_i (\hat{a}^1_1 \hat{a}^1_i - \hat{a}^2_1 \hat{a}^2_i)^* - \hat{C} \right),
\]
we have \( \mathbb{E}(\hat{X}_i) = 0 \), and \( \| \hat{X}_i \| \leq \frac{4M}{m} \). Moreover \( \mathbb{E}(\hat{X}_i^2) \leq \frac{2M}{m}\hat{C} \). Hence we have: \( \mathbb{E} \left( \sum_i \hat{X}_i \right)^2 \leq \frac{2M}{m}\hat{C} \).

The intrinsic dimension \( \text{intdim}(\hat{C}) = \frac{\text{tr}(\hat{C})}{\|\hat{C}\|} = \frac{1}{\hat{\lambda}} = 1 \). We are now ready to apply Theorem 9:

\[
\mathbb{P} \left( \left\| \hat{C}_m - \hat{C} \right\| \geq t \right) \leq 4\text{intdim}(\hat{C}) \exp \left( \frac{-t^2/2}{\frac{cM}{m}\|\hat{C}\| + \frac{4M}{m}t^2/3} \right),
\]
is valid for sufficiently large \( t \). To achieve a relative error \( \epsilon \), such that \( 0 < \epsilon < 1 \), we have for

\[
m \geq \frac{cM}{\epsilon^2\|\hat{C}\|}, \quad \left\| \hat{C}_m - \hat{C} \right\| \leq \epsilon\|\hat{C}\|,
\]
with probability \( 1 - \text{intdim}(\hat{C})^c = 1 - \epsilon' \).

**Putting all together:**

Setting \( \beta = \sqrt{2} \). We have with probability at least \( 1 - 2me^{-2n} - \epsilon' \), for \( m \geq \frac{cM}{\epsilon^2\|\hat{C}\|} \)

\[
\left\| \hat{C}_m - C \right\| \leq \epsilon\|\hat{C}\| + 2e^{-n}
\]

where \( M = 2n(1 + \sqrt{2})^2 \). Therefore for sufficiently large \( n \), for \( m \geq \frac{cM}{\lambda \epsilon^2}, \left\| \hat{C}_m - C \right\| \leq \epsilon\lambda \) with probability at least \( 1 - O(me^{-2n}) \).
The proof of Theorem 8 follows easily from Lemma 11 and Proposition 8:

**Proof of Theorem 8.** By Lemma 11, and Proposition 8 we conclude for sufficiently large $n$: For $m \geq \frac{\epsilon}{\lambda^2}$, $||\tilde{x}_m \tilde{x}_m^* - x_0 x_0^*||_F \leq \epsilon$ with probability at least $1 - O(me^{-2n})$, where $c$ is a sufficiently large numeric constant.

### 9.9.1 Alternative Proof

We state here an alternative proof to Proposition 4. The technique is different and is of independent interest.

**Proof of Proposition 4.** For $x \in \mathbb{C}^n$, $||x|| = 1$.

\[
\mathcal{E}^{x_0}(x) = \mathbb{E} \left( y \left( |\langle a^1, x \rangle|^2 - |\langle a^2, x \rangle|^2 \right) \right),
\]

(9.34)

where $y = \text{sign}(|\langle a^1, x_0 \rangle|^2 - |\langle a^2, x_0 \rangle|^2)$. Recall $a^1, a^2 \sim \mathcal{CN}(0, I_n)$ are complex Gaussian vectors, there exists $g, h \sim \mathcal{N}(0, \frac{1}{2}) + i\mathcal{N}(0, \frac{1}{2})$ i.i.d. and $G, H \sim \mathcal{N}(0, \frac{1}{2}) + i\mathcal{N}(0, \frac{1}{2})$ i.i.d.,

\[
\begin{align*}
\langle a^1, x_0 \rangle &= g, \\
\langle a^2, x_0 \rangle &= G,
\end{align*}
\]

\[
\begin{align*}
\langle a^1, x \rangle &= \langle x_0, x \rangle g + \sqrt{1 - |\langle x_0, x \rangle|^2} h. \\
\langle a^2, x \rangle &= \langle x_0, x \rangle G + \sqrt{1 - |\langle x_0, x \rangle|^2} H.
\end{align*}
\]

\[
|\langle a^1, x \rangle|^2 - |\langle a^2, x \rangle|^2 = \left| \langle x_0, x \rangle g + \sqrt{1 - |\langle x_0, x \rangle|^2} h \right|^2 - \left| \langle x_0, x \rangle G + \sqrt{1 - |\langle x_0, x \rangle|^2} H \right|^2
\]

\[
= |\langle x_0, x \rangle|^2 (|g|^2 - |G|^2) + (1 - |\langle x_0, x \rangle|^2) (|h|^2 - |H|^2) + 2\Re \left( \langle x_0, x \rangle \sqrt{1 - |\langle x_0, x \rangle|^2} (\bar{g}h - \bar{G}H) \right),
\]

where $\Re(z)$ denotes the real part of $z \in \mathbb{C}$. Recall that $y = \text{sign}(\theta(|g|^2) - \theta(|G|^2))$. From Lemma 13, we know that $E_1 = |g|^2$, and $E_2 = |G|^2$ are two exponential independent random variables $\text{Exp}(1)$.

Note that the random variables $y = \text{sign}(\theta(|g|^2) - \theta(|G|^2))$, and $(|h|^2 - |H|^2)$ are independent, and hence $\mathbb{E} (y(|h|^2 - |H|^2)) = 0$. Moreover $\mathbb{E} (\langle x \rangle) = \mathbb{E}_{g,G} (\langle \bar{g}h - \bar{G}H \rangle) = 0$. Hence, we have:

\[
\mathcal{E}^{x_0}(x) = \mathbb{E} \left( y \left( |\langle a^1, x \rangle|^2 - |\langle a^2, x \rangle|^2 \right) \right)
\]

\[
= |\langle x_0, x \rangle|^2 \mathbb{E} (\text{sign}(\theta(E_1) - \theta(E_2))(E_1 - E_2))
\]

\[
= \lambda |\langle x_0, x \rangle|^2.
\]

### 9.9.2 Technical Tools

Here, we collect a few technical results needed in the proofs.
Lemma 13 (Spacing of Exponentials). For \( x_0 \in \mathbb{C}^n \), and \( a \sim \mathcal{N}(0, \frac{1}{2}I_n) + i\mathcal{N}(0, \frac{1}{2}I_n) \), 
\( E = |\langle a, x_0 \rangle|^2 \) follows an exponential distribution with parameter one (see [Mro10] for a proof).

Moreover [SW], if we let and \( a^1, a^2 \) i.i.d. \( \sim \mathcal{N}(0, \frac{1}{2}I_n) + i\mathcal{N}(0, \frac{1}{2}I_n) \), let \( E_1(x_0) = |\langle a^1, x_0 \rangle|^2 \), and \( E_2(x_0) = |\langle a^2, x_0 \rangle|^2 \), and \( E^{(1)}(x_0) \) and \( E^{(2)}(x_0) \), then the corresponding order statistics \( \text{i.e } E^{(2)}(x_0) \geq E^{(1)}(x_0) \). \( \Delta(x_0, x_0) = |\langle a^{(2)}, x_0 \rangle|^2 - |\langle a^{(1)}, x_0 \rangle|^2 \) is also exponentially distributed with parameter one. \( \Delta \) is called spacing of order statistics of exponentials.

Lemma 14 (Laplace Exponential). \( U \sim \exp(\gamma) \), \( V \sim \exp(-\gamma) \), \( U \) and \( V \) are independent then \( U - V \sim \text{Laplace}(0, \frac{1}{\gamma}) \). The CDF of \( U - V \) is :

\[
F_{U-V}(z) = \frac{1}{2} + \frac{1}{2} \text{sign}(z)(1 - \exp(-\gamma|z|)).
\]

Theorem 10 (Non commutative Bernstein Inequality [Ver11]). Consider a finite sequence \( X_i \) of independent centered self adjoint random \( n \times n \) matrices. Assume we have for some numbers \( K \) and \( \sigma \) that:

\[
||X_i|| \leq K \text{ almost surely } \quad \left\| \sum_i E X_i^2 \right\| \leq \sigma^2.
\]

Then, for every \( t > 0 \), we have:

\[
P \left\{ \left\| \sum_i X_i \right\| > t \right\} \leq 2n \exp \left( \frac{-t^2/2}{\sigma^2 + Kt/3} \right).
\]

Theorem 11 (Covariance Estimation for Arbitrary distributions [Ver11]). Consider a distribution with covariance matrix \( \Sigma \) supposed in some centered ball whose radius we denote \( \sqrt{M} \). Let \( \Sigma_m \) be the empirical covariance. Let \( \epsilon \in ]0, 1[ \), and \( t \geq 1 \). Then the following holds with probability at least \( 1 - n^{-t^2} \):

If \( m \geq C(t/\epsilon)^2||\Sigma||^{-1}M\log(n) \) then \( ||\Sigma_m - \Sigma|| \leq \epsilon||\Sigma||. \)

135
Chapter 10

One Bit Coded Diffraction Patterns: Phase Retrieval and Super-Resolution

In this chapter we study a realistic setup for phase retrieval, where the signal of interest is modulated or masked and then for each modulation or mask a diffraction pattern is collected, producing a coded diffraction pattern (CDP) [CLM13]. We are interested in the setup where the resolution of the collected CDP is limited by the Fraunhofer diffraction limit of the imaging system. We investigate a novel approach based on a geometric quantization scheme of phase-less linear measurements into (one-bit) coded diffraction patterns, and a corresponding recovery scheme. The key novelty in this approach consists in comparing pairs of coded diffractions patterns across frequencies: the one bit measurements obtained rely on the order statistics of the un-quantized measurements rather than their values. This results in a robust phase recovery, and unlike currently available methods, allows to efficiently perform phase recovery from measurements affected by severe (possibly unknown) non linear, rank preserving perturbations, such as distortions. Another important feature of this approach consists in the fact that it enables also super-resolution and blind-deconvolution, beyond the diffraction limit of a given imaging system.

10.1 Introduction

10.1.1 The Phase Retrieval Problem and the Diffraction Limit

The problem of phase retrieval is ubiquitous in many areas of imaging science and engineering, where we are able to measure only magnitude of measurements. The phase recovery problem can be modeled as the problem of reconstructing a $n$-dimensional complex vector $x_0$ given only the magnitude of $m$ phase-less linear measurements. Such a problem arises for example in X-ray crystallography [Har93, Lea08], diffraction imaging [BDP+07, Rod08] or microscopy [MISE08], where one can only measure the intensities of the incoming waves, and wishes to recover the lost phase in order to be able to reconstruct the desired object.
Formally speaking for a given vector $x_0 \in \mathbb{C}^n$ (without loss of generality we assume $n$ to be even), we wish to measure $\langle a_k, x \rangle$, but the only available information is of the form:

$$b_k = \theta(|\langle a_k, x \rangle|^2), \; k = 1 \ldots m,$$

(10.1)

where $a_k$ is a set of sampling vector in $\mathbb{C}^n$, and $\theta$ models possibly unknown non-linear perturbations of the values: distortion and exponential noise for instance.

Recovering signals from the modulus of their Fourier transform is at the core of the phase retrieval problem. For instance in coherent X-ray crystallography [Lea08], speckle imaging in astronomy [Fri66], in microscopy [Gus00] or more broadly in Fourier optics it follows from the Fraunhofer diffraction principle that the optical field at the detector can be approximated by the Fourier transform of the sensed object. Since light detectors can measure only intensities of the incoming waves the problem is therefore to recover the discrete signal $x_0 \in \mathbb{C}^n$ from measurements of the type:

$$b_k = \theta \left( \left| \sum_{j=1}^{n} x_0[j] e^{-i2\pi(j-1)\frac{(k-1)}{n}} \right|^2 \right), \; k \in \Omega, \; \Omega \subseteq [1, n]$$

(10.2)

where $\Omega$ represents a set of sampled frequencies, and $\theta$ a possibly unknown non-linearity. When compared to (10.1) we note that in (10.2) $a_k$ correspond to a set of sampled complex sinusoids. When $\Omega = [1, n]$, we have the full knowledge of the modulus of the Fourier series decomposition of the signal of interest on whole ranges of frequencies. In practice due to the Fraunhofer diffraction limit we are able to measure intensities of the Fourier transform within a range of frequencies below the so called cut-off frequency $f_c$. Hence the information we have available about $x_0$ is a sample of the lower end of its power spectra in the form of the lowest $2f_c + 1$ modulus of the Fourier series coefficients ($f_c$ is an integer).

For instance in microscopy imaging with coherent illumination the object of interest is diffracted through a lens. The lens is characterized by its Point Spread Function (PSF) $h$, and the cut-off frequency $f_c = \frac{2\pi NA}{\nu}$, where $NA$ is the numerical aperture of the lens and $\nu$ is the wavelength of the illumination light. Let $\hat{h}, \hat{x}_0$ be the Fourier transform of the PSF and the signal $x_0$ respectively, for continuous signals we know that:

$$\hat{h}(w) = 0 \quad \text{for} |w| > f_c.$$

Hence we measure the modulus of the Fourier transform of $x_0 \ast h$ rather than $x_0$, where $\ast$ denotes the discrete convolution operation. If we set $\theta = \theta_h$, where

$$\theta_h \left( \left| \sum_{j=1}^{n} x_0[j] e^{-i2\pi(j-1)\frac{(k-1)}{n}} \right|^2 \right) = \left| \sum_{j=1}^{n} (x_0 \ast h)[j] e^{-i2\pi(j-1)\frac{(k-1)}{n}} \right|^2,$$
equation (10.2) becomes (with some abuse of notation, in re-indexing $k$ between $-\frac{n}{2}$ and $\frac{n}{2} - 1$):

$$b_k = |h_k|^2 |\hat{x}_0|^2_k \quad k \in [-f_c, f_c] \text{ and } b_k = 0 \text{ elsewhere.}$$

(10.3)

It follows from equation (10.3), that in addition to the missing phase problem we are facing a super-resolution problem since the high frequency content of the signal is also lost due to the physical resolution limit induced by the cut-off frequency $f_c$ (See for e.g. [Gus00]).

This chapter answers the following questions:

1. **Robust recovery:** Is it possible to robustly recover the missing phase from the power spectra of a signal $x_0$ that is undergoing severe unknown non linear distortions or a stochastic noise?

2. **Phase Recovery, Super-Resolution, and Blind deconvolution:** Is it possible to recover the signal from the lower end of its power spectra? In other words is it possible to super-resolve the signal beyond the diffraction limit of a given optical system even if the PSF of that system was unknown (blind deconvolution)?

**Notations:** $\ast$ represents a convolution, and $\odot$ the Hadamard product (component-wise product). For $z \in \mathbb{C}$, $|z|^2$ is squared complex modulus of $z$. For $a, a' \in \mathbb{C}^n$, $\langle a, a' \rangle$ is the complex dot product in $\mathbb{C}^n$. For $a \in \mathbb{C}^n$, $a^*$ is the complex conjugate of $a$, $||a||_2$ is the norm 2 of $a$ and $\cdot$. Let $A$ a complex hermitian matrix in $\mathbb{C}^n$, $||A||_F$ denotes the Frobenius norm of $A$, $||A||$ denotes the operator norm of $A$, $\text{Tr}(A)$ denotes the trace of $A$. Throughout the chapter, we denote by $c, C$ positive absolute constants whose values may change from instance to instance.

### 10.1.2 Phase Retrieval from Coded Diffraction Patterns (CDP):

**Previous Work**

As mentioned in the introduction the set of sampling vectors we are interested in, is the set of complex sinusoids. While the Gaussian measurements setting discussed in Chapter 9 allows to carry an interesting theory and gives a glimpse on the efficiency of proposed methods in more practical setups, it is of great interest to study the Fourier sampling mentioned in the introduction. A practical setup consists in modulating the signal with multiple structured illuminations for instance, and then measuring multiple diffraction patterns of the modulated signals. The modulation step could be replaced by masking the signal of interest with an appropriate mask. This is indeed an attractive framework to resolve the ambiguity in the phase retrieval problem. Firstly suggested in [Mis73], this technique comes under different names: digital holography [YZ97], ptychography [HKH+13], Fourier ptychographic microscopy [ZHY13], etc.. These techniques yield to many successful applications in structured illumination microscopy [Gus00] and more broadly in many linear and non linear Fourier optics applications, where both phase retrieval and super-resolution are achieved via masking or the use of multiple structured illumination modulation. Let $w_\ell \in \mathbb{C}^n, \ell = 1 \ldots r$
be the modulating waves (or the masks), we observe the following coded diffraction patterns [CLM13]:

\[
    b_{\ell,k} = \theta \left( \left| \sum_{j=1}^{n} x_0[j] w_{\ell}[j] e^{-i2\pi(j-1)(k-1)/n} \right|^2 \right), \quad k \in \Omega, \quad \Omega \subseteq [1,n], \quad \ell = 1 \ldots r. \quad (10.4)
\]

In other words, noting \( F \) the Discrete Fourier Transform (DFT) Matrix, \( Diag(w) \) the diagonal matrix with the modulation pattern on its diagonal, and \( \Omega = [1,n] \) we have:

\[
    b_{\ell} = \theta(|F Diag(w) x_0|)^2 \in \mathbb{R}_+^n, \quad \ell = 1 \ldots r, \quad (10.5)
\]

where \( \theta \) and the complex modulus act component-wise.

### 10.1.3 Convex Approach: PhaseLift

In the setting of Coded diffraction pattern, in a recent work, for \( \theta(z) = z \), and a set of admissible modulations authors in [CLM13] show that an approach similar to Phase-lift allows the exact recovery of the signal with high probability \((1 - n^{-\gamma})\) given that:

\[
    r \geq c\gamma \log^4 n,
\]

for a fixed numerical constant \( c \). This result was improved in [GKK14] to

\[
    r \geq c\omega \log^2 n
\]

with a probability guarantee of \( 1 - e^{-\omega} \). Hence the number of modulations is of \( O(\gamma \log^3 n) \) with a probability guarantee of \( 1 - n^{-\gamma} \).

### 10.1.4 Non Convex Approaches: Alternating Minimzation , Wirtinger Flow, Polarization

While powerful, the convex relaxation approach incur in cumbersome computations, and in practice non convex approaches based on greedy alternating minimization (AM) [GS72, Fie82, GL84] are often used. The analysis of Alternating minimization for coded diffraction patterns has not been addressed yet in the literature and seems challenging technically. More recently, an approach based on suitably initialized gradient descend via wirtinger flow, for CDP was introduced in [CLS14]. This approach parallels the one for Gaussian measurements introduced in the same paper by [CLS14]. The initialization step is similar to subexpPhase, and consists in finding the maximum eigenvector \( \hat{x}_m \) of the following empirical matrix \( \hat{C}_r \).
(suitably normalized as discussed in Chapter 9):

\[
\hat{\mathbf{C}}_r = \frac{1}{r} \sum_{i=1}^{r} \text{Diag}(w_i) F \text{Diag}(b_i) F^* \text{Diag}(w_i^*),
\]

(10.6)

The rest of the updates follow the same procedure of the gaussian measurement case, via a resampled Wirtinger flow (See Equation 9.14). It is shown in [CLS14], that for:

\[
r > c_0 \log^4 n, \text{ dist}(z_r, x_0) \leq \frac{1}{8\sqrt{n}} \left(1 - \frac{\mu}{3}\right)^{r/2} ||x_0||,
\]

with a probability at least \(1 - (2r + 1)/n^3 + 1/n^2\).

Finally in [ABFM12, ASBM13] authors introduce another approach to phase retrieval by polarization. In [ASBM13] authors propose a construction of binary masks that ensures phase recovery by polarization. It is shown in [ASBM13] that \(O(\log(n))\) binary masks are needed to ensure recovery in the noiseless case.

### 10.1.5 This Chapter: One Bit Coded Diffraction Patterns

In this chapter we are interested in the setting where \(\theta\) is different from the identity. We restrict our analysis to complex Gaussian modulations. Three settings are of interest:

1. **Noise:**
   - *Additive Stochastic noise: We observe noisy coded diffraction patterns,*
     \[
     b_\ell = |F \text{Diag}(w_\ell) x_0|^2 + \nu_\ell \in \mathbb{R}_+^n \quad \ell = 1 \ldots 2r,
     \]
     (10.7)
     where \(\nu_\ell\) are independent exponential vectors \(\text{Exp}(\gamma), (\sigma = \frac{1}{\gamma})\).
   - *Poisson Noise: We observe noisy coded diffraction patterns contaminated with poisson noise,*
     \[
     b_\ell = \mathcal{P}_\eta(|F \text{Diag}(w_\ell) x_0|^2) \in \mathbb{R}_+^n \quad \ell = 1 \ldots 2r,
     \]
     (10.8)
     where \(\mathcal{P}_\eta\) is a component-wise poisson noise : For \(z, \eta > 0, \mathcal{P}_\eta(z) \sim \text{Poisson}(\frac{z}{\eta})\).

2. **Distortion:** We observe distorted coded diffraction patterns:

   \[
   b_\ell = \tanh (\alpha |F \text{Diag}(w_\ell) x_0|^2) \in \mathbb{R}_+^n \quad \ell = 1 \ldots 2r, \alpha > 0
   \]
   (with some abuse of notations \(\tanh\) acts component-wise).

3. **Diffraction Limit/Super-Resolution/Blind deconvolution:** The modulated signal diffracts through a lens characterized by a PSF \(h\) and a cut-off frequency \(f_c\). Let \(H\) be the Toeplitz
In this chapter we take the point of view of [MR13] and define a quantization scheme for the coded diffraction patterns, by comparing pairs of coded diffraction patterns. Consider pairs of coded diffraction patterns associated to pairs of independent modulations \((w_1, w_2)\), where \(w_1, w_2 \sim \mathcal{CN}(0, I_n)\):

\[
(b_1^i, b_2^i) = (\theta(|FD\text{Iag}(w_1^i)x_0|^2), \theta(|FD\text{Iag}(w_2^i)x_0|^2)) \in \mathbb{R}_+^r \times \mathbb{R}_+^r, \quad i = 1 \ldots r. \quad (10.11)
\]

For each pair \((b_1^i, b_2^i)\) of coded diffraction patterns we define a one bit coded diffraction pattern as:

\[
y_i \in \{-1, 1\}^n, \quad y_i = \text{sign}(b_1^i - b_2^i), \quad i = 1 \ldots r. \quad (10.12)
\]

Now the One Bit Phase Retrieval problem consists in finding \(x_0\) from the knowledge of one bit coded diffraction patterns \((y_1 \ldots y_r)\). Similarly to One bit phase retrieval from Gaussian measurements we show that the phase retrieval problem from one bit coded diffraction patterns reduces to finding the maximum eigen-vector \(\hat{x}_r\) of the matrix \(\hat{C}_r\):

\[
\hat{C}_r = \frac{1}{r} \sum_{i=1}^r (\text{Diag}(w_1^i)FD\text{Iag}(y_i)F^*\text{Diag}(w_1^i) - \text{Diag}(w_2^i)FD\text{Iag}(y_i)F^*\text{Diag}(w_2^i)),
\]

Indeed in this chapter we show that:

\[
\mathbb{E}(\hat{C}_r) = \lambda x_0 x_0^*,
\]

where \(\lambda\) depends on \(\theta\).

### 10.2 Main Results

In the following we give the only assumption we make on \(\theta\) throughout the chapter, and state our main results for the three setups of interest discussed in Section 10.1.5. As mentioned before, we assume that \(\theta\) preserves the ranking of the intensities. We shall make one assumption on the non linearity \(\theta\),

\[
\lambda = \mathbb{E}(\langle \text{sign}(\theta(E_1) - \theta(E_2)), (E_1 - E_2) \rangle) > 0, \quad (10.14)
\]

where \(E_1, E_2\) are two independently distributed exponential random \(n\)-dimensional vectors with mean \(\frac{1}{n}\). To see why this assumption is natural, notice that \(|FD\text{Iag}(w)x_0|^2 \sim \mathcal{CN}(0, I_n)|w|^2\).

141
\( (\frac{1}{n} \text{Exp}(1))^{\otimes n} \) if \( w \sim \mathcal{CN}(0, I_n) \) and \( ||x_0|| = 1 \), thus

\[
\mathbb{E}(\langle y_1, |F \text{Diag}(w_1)x_0|^2 - |F \text{Diag}(w_2)x_0|^2 \rangle) = \mathbb{E}(\langle \text{sign}(\theta(E_1) - \theta(E_2)), (E_1 - E_2) \rangle) = \lambda > 0.
\]

Then the above assumption simply means that the one bit measurements preserve robustly the ranking of the intensities. Let \( \hat{x}_r \) be the maximum eigenvector of \( \hat{C}_r \) defined in (10.13). The following Theorem shows that \( \hat{x}_r \) is an \( \epsilon^- \) estimate of \( x_0 \).

**Theorem 12 (Phase Retrieval From One Bit Coded Diffraction Patterns).** For \( x_0 \in \mathbb{C}^n, ||x_0|| = 1, \) and \( 0 < \epsilon < 1 \). Assume \( y_1 \ldots y_r, \) follow the model given in (10.12). Then we have with a probability at least \( 1 - O(n^{-2}) \),

\[
\text{for } r \geq \frac{c}{\epsilon^2 \lambda^2} \log^3 n, \quad ||\hat{x}_r - x_0 e^{i\phi}||^2 \leq \epsilon
\]

where \( c \) is a numeric constant, and \( \phi \in [0, 2\pi] \) is a global phase. \( \lambda \) is given in (10.14).

For the noiseless model \( \theta(z) = z \) and \( \lambda = 1 \). Thus the theorem states that \( O(\log^3 n) \) pairs of coded diffraction patterns ensures the recovery of the phase. For different observation model \( \theta \) it suffices to compute the value of \( \lambda \) as given in (9.28). We turn now to the noisy measurements setup (10.7) and show robustness of phase retrieval from one bit coded diffraction patterns:

**Corollary 3 (One bit Recovery/ Noise).** For \( x_0 \in \mathbb{C}^n, ||x_0|| = 1, \) and \( 0 < \epsilon < 1 \). Assume \( y_1 \ldots y_r, \) follow the model given in (10.12), for \( \theta(z) = z + \nu, \nu \sim \text{Exp}(\gamma) \). Where \( \nu \) is an exponential noise with variance \( \sigma = \frac{1}{\epsilon^2} \). Then for any \( \epsilon, 0 < \epsilon < 1, \) we have with a probability at least \( 1 - O(n^{-2}) \),

\[
\text{for } r \geq \frac{c}{\epsilon^2} (1 + \sqrt{\sigma})^4 \log^3 n, \quad ||\hat{x}_r - x_0 e^{i\phi}||^2 \leq \epsilon,
\]

where \( c \) is a numeric constant, and \( \phi \in [0, 2\pi] \) is a global phase.

In other words, under an exponential noise we have:

\[
||\hat{x}_r - x_0 e^{i\phi}||_2 \leq C \sqrt{\frac{\log^3 n (1 + \sqrt{\sigma})^2}{r (1 + 2\sqrt{\sigma})}}.
\]

Beyond robustness to noise, another desirable feature for phase retrieval from phase-less measurements, is the robustness to distortions of the values of intensities. Is it possible to retrieve the phase from coded diffraction patterns that are undergoing clipping for instance (as in equation (10.9))? 

**Corollary 4 (One bit Recovery/ Distortion).** For \( x_0 \in \mathbb{C}^n, ||x_0|| = 1, \) and \( \epsilon > 0 \). Assume \( y_1 \ldots y_m, \) follow the model given in (10.12), for \( \theta(z) = \text{tanh}(\alpha z), \alpha > 0 \). Then for any
\(\epsilon, 0 < \epsilon < 1,\) we have with a probability at least \(1 - O(n^{-2})\),

\[
\text{for } r \geq \frac{c \log^3 n}{\epsilon^2 \lambda^2(\alpha)}, \quad ||\hat{x_r} - x_0 e^{i\phi}||^2 \leq \epsilon,
\]

where \(c\) is a numeric constant, and \(\phi \in [0, 2\pi]\) is a global phase. \(\lambda(\alpha) = \mathbb{E}(|E_1 - E_2| \text{sign}(1 - \tanh(\alpha E_1) \tanh(\alpha E_2)))\), is a decreasing function in \(\alpha\).

For the last setup where the resolution of the observed diffraction patterns is limited by the Fraunhofer diffraction limit \(f_c\) of an optical system as in equation (10.10). We show that the recovery is still possible even if the PSF of the optical system was unknown. The number of modulations needed is poly-logarithmic in the dimension and quadratic in the super-resolution factor \(\text{SRF} = \frac{n}{2f_c + 1}\) defined in [FGC12].

**Corollary 5** (One bit Recovery/Super-Resolution). For \(x_0 \in \mathbb{C}^n, ||x_0|| = 1,\) and \(0 < \epsilon < 1.\) Assume \(y_1 \ldots y_m,\) follow the model given in (10.12) for \((b_1^1, b_1^2)\) defined as in (10.10) for a PSF \(h\) characterized by the cut-off frequency \(f_c.\) Then for \(\epsilon, 0 < \epsilon < 1,\) we have with we have with a probability at least \(1 - O(n^{-2})\),

\[
\text{for } r \geq \frac{c}{\epsilon^2} (\text{SRF})^2 \log^3 n, \quad ||\hat{x_r} - x_0 e^{i\phi}||^2 \leq \epsilon.
\]

where \(c\) is numeric constant, and \(\phi \in [0, 2\pi]\) is a global phase. \(\text{SRF}\) is the super-Resolution factor defined as: \(\text{SRF} = \frac{n}{2f_c + 1}.\)

It follows that:

\[r \geq \frac{c}{\epsilon^2} \log^3 n \quad ||\hat{x_r} - x_0 e^{i\phi}||^2 \leq \text{SRF} \epsilon,\]

this dependency on the Super-Resolution Factor (SRF) is similar to results in [FGC12], where super-resolution is achieved via total variation norm minimization and linear measurements. Note that in [FGC12] the phase of the linear measurements, and the PSF \(h\) are assumed to be known.

It is worth noting that in [FGC12] the super-resolution problem considered, is different from our setting as authors consider a harder problem: super-resolution from a single image, and a strong prior, namely a point sources model and a total variation norm minimization. In our case we have access to multiple coded diffraction patterns and this is known as multi-frame super-resolution see for example [Gus00] and references therein.

The proof of Corollary 5 is given in Section 10.4. Corollary 5 states a surprising fact: one bit coded diffraction patterns allow not only the super-resolution of the signal but also it leads to a blind deconvolution since the only information needed on \(h\) is its super-resolution factor \(\text{SRF},\) its PSF might be completely unknown. Intuitively the random modulations push the high frequency content of \(x_0\) to the frequency interval where the Fourier transform of \(h\) is non zero. The number of modulations needed is therefore naturally proportional to the \(\text{SRF}\) as shown in Corollary 5. Hence the high frequency content of \(x_0\) is mapped to the lower end spectrum by modulation or masking. Phase retrieval from one bit coded diffraction patterns
in a way estimates the missing phase, the missing high frequency content and corrects for the blur induced by the unknown PSF $h$.

Moreover this result is still true if the observation model was:

$$(b_1^i, b_2^i) = (|FH_i\text{Diag}(w_1^i)x_0|^2, |FH_i\text{Diag}(w_2^i)x_0|^2) \quad i = 1 \ldots r,$$

$$y_i \in \{-1, 1\}^n, \quad y_i = \text{sign}(b_1^i - b_2^i), \quad i = 1 \ldots r,$$

(10.15)

where $H_i$ are Toeplitz matrices associated to different unknown stochastic perturbations $h_i$. We assume for simplicity that the Fourier transform of $h_i$ are non zero in the same frequency domain (the result is still true if this was not the case we don’t analyze this case in this chapter). The only requirement is therefore to have the same perturbation on each considered pairs of coded diffraction patterns. For example in microscopy small perturbations will result in a change in the PSF. In astronomy in speckle imaging different $h_i$ model different atmospheric perturbations in a long exposure acquisition.

Surprisingly one bit coded diffraction patterns allow blind deconvolution even in the case of varying PSFs.

### 10.2.1 Discussion and Perspectives

**Discussion**

We comment in this section on our results and compare them to the current state of the art and put them in the perspective of future research. Let $f_j$ be a row of the DFT matrix.

For phase-lift, by inspecting the proof in [CLM13] we note that three factors govern the sample complexity $O(\log^4(n))$, the first two of them come from matrix concentration inequality and the last one is due to the golfing scheme:

- A bound on the measurements, $|f_j\text{Diag}(w)x_0|^2 j = 1 \ldots m$: $|f_j\text{Diag}(w)x_0|^2 \leq \beta \log(n)$ with high probability.

- A bound on the the absolute values of the entries of the modulation $|w_i|^2, i = 1 \ldots n$.

  In [CLM13], authors define a family of admissible modulation, such that among other conditions: $|w_i|^2 \leq M$, where $M$ is a constant independent to the dimension. It is worth noting that this class of modulations as opposed to a complex Gaussian modulation, saves extra poly-logarithmic terms in the overall sample complexity of that approach.

- An extra $\log(n)$ in the sample complexity is needed for the golfing scheme.

In contrast in our case the saving of extra poly-logarithmic terms ($O(\log^3(n))$) comes from the nature of one bit coded diffraction patterns. Our one bit measurements are bounded by one, hence they do not contribute to the sample complexity. On the other hand our modulations are Complex Gaussian. Complex Gaussian modulations have their squared absolute values bounded with high probability $|w_i|^2 \leq \beta \log(n)$ and hence they contribute to the sample complexity.
complexity.

Perspectives

Greedy Refinements. Indeed PhaseLift and one bit phase retrieval are not comparable since one achieves exact recovery and the other achieves approximate recovery. For an accuracy $\epsilon$ the sample complexity for one bit phase retrieval scales as $\frac{1}{\epsilon^2}$. As for the Gaussian case the Alternating minimization [NJS13] for coded diffraction patterns initialized with the one bit solution would guarantee a better dependency on $\epsilon$, we leave that direction to a future research. We conjecture that $O(\log^2 n + \log \frac{1}{\epsilon} \log \log \frac{1}{\epsilon})$ pairs of coded diffraction patterns ensures $\epsilon$ recovery with the alternating minimization initialized with the one bit solution. Our experiments on both simulated data and images confirm that (See Section 10.6). The wirtinger flow approach could be also initialized with one bit solution, the analysis is independent to the initialization, hence we can get similar results to the one in [CLS14].

Non Gaussian Modulations or Masks. Another direction would be to investigate admissible modulation of [CLM13] and one bit measurements as they both enjoy dimensionless boundedness, we leave also that point to a future work.

Blind Deconvolution from Coded Diffraction Patterns. In microscopy the PSF of the lens is often known. One Bit solution is agnostic to the PSF, hence one bit phase retrieval offers a good initial point to the alternating minimization conditioned on the knowledge of the PSF: phase retrieval with blur correction (See Section 10.6). For an unknown PSF an open question remains on how to provably recover both the signal and the PSF, via alternating minimization suitably initialized.

10.2.2 Roadmap

The chapter is organized as follows: In Section 10.3 we introduce the one bit coded diffraction patterns scheme and the corresponding phase recovery procedure. In section 10.4 we show how super-resolution can be tackled within our framework. We address algorithms and computational aspects in Sections 10.5 and 10.6. Finally we give the proofs in Section 10.7.

10.3 Quantizing Coded Diffraction Patterns

10.3.1 Preliminary Matrix Notation

Let $M \in \mathbb{C}^{n \times n}$ be a complex matrix, $\text{diag}(M)$ is a vector in $\mathbb{C}^n$, containing the diagonal elements of $M$.

Let $u \in \mathbb{C}^n$ be a complex vector, $\text{Diag}(u)$ is a matrix in $\mathbb{C}^{n \times n}$, with $u$ on the diagonal and
zeros elsewhere. 
Let $F$ be the discrete Fourier matrix, such that $F_{jk} = \frac{1}{\sqrt{n}} e^{-\frac{2\pi i (j-1)(k-1)}{n}}$, $j = 1 \ldots n, k = 1 \ldots n$.

### 10.3.2 One Bit Coded Diffraction Patterns

We start by defining the quantization scheme of the values of masked Fourier intensities or CDP. We assume that we observe $\theta(\lvert F\text{Diag}(w^1)x_0\rvert^2)$, where $\theta$ is eventually an unknown non-linearity satisfying (9.28). Following the same procedure in the Gaussian case we quantize the differential of two independent coded diffraction patterns.

**Definition 22** (One-bit Fourier quantizer). Let $W = (w^1, w^2)$, where $w^1, w^2$ are i.i.d. complex Gaussian vectors $\mathcal{N}(0, \frac{1}{2}I_n) + i\mathcal{N}(0, \frac{1}{2}I_n)$. $w^1$ and $w^2$ are called Gaussian masks or modulations. Let $F$ be the discrete Fourier matrix in $\mathbb{C}^{n \times n}$. For $x_0 \in \mathbb{C}^n$, a one bit quantizer of coded diffraction patterns is given by

\[ Q_{W}^\theta : \mathbb{C}^n \to \{-1, 1\}^n, \quad Q_{A}^\theta(x_0) = \text{sign} \left( \theta(\lvert F\text{Diag}(w^1)x_0\rvert^2) - \theta(\lvert F\text{Diag}(w^2)x_0\rvert^2) \right). \]

where $\lvert F\text{Diag}(w^1)x_0\rvert^2$ is the complex modulus of each component of $F\text{Diag}(w^1)x_0$. $\theta$ is the observation model. $\theta$ is eventually an unknown non-linearity that satisfies equation (9.28).

Recall that a basic quantizer in the noiseless case is obtained setting $\theta(z) = z$.

Now for a total of $2r$ masks or modulations we define the one bit coded diffraction patterns:

**Definition 23** (One Bit Coded Diffraction Patterns). Let $\{W_i = (w^1_i, w^2_i)\}_{1 \leq i \leq r}$, be $2r$ i.i.d. Gaussian masks in $\mathbb{C}^n$, and $Q_{W_i}^\theta(x_0)$ as in Def 22. The Quantized Phase-less sensing is : $Q : \mathbb{C}^n \to \{-1, 1\}^{nr}$, $Q(x_0) = (Q_{W_1}^\theta(x_0), \ldots, Q_{W_n}^\theta(x_0))$.

Let

\[ y_i = \text{sign} \left( \theta(\lvert F\text{Diag}(w^1_i)x\rvert^2) - \theta(\lvert F\text{Diag}(w^2_i)x\rvert^2) \right) \in \{-1, 1\}^n \quad i = 1 \ldots r \quad (10.16) \]

In this chapter, we are interested in recovering $x_0$ from its one bit coded diffraction patterns $y = (y_1 \ldots y_m) = Q(x_0)$. It is easy to see that the phase retrieval amounts to the following feasibility problem:

\[
\text{find } x \\
\text{subject to } \langle y_i, |F\text{Diag}(w^1_i)x|^2 - |F\text{Diag}(w^2_i)x|^2 \rangle \geq 0, \quad i = 1 \ldots r. \\
||x||^2 = 1. \quad (10.17)
\]

Again we propose the following relaxation to tackle that problem:

\[
\max_{x, ||x||^2 = 1} \left( \frac{1}{r} \sum_{i=1}^{r} \langle y_i, |F\text{Diag}(w^1_i)x|^2 - |F\text{Diag}(w^2_i)x|^2 \rangle \right) \quad (10.18)
\]
The proof architecture is similar to the Gaussian case. Proofs are given in Section 10.7. We start by a preliminary definition:

**Definition 24** (Fourier Risk and Empirical risk). Let $x_0 \in \mathbb{C}^n$, $||x_0|| = 1$. For $x \in \mathbb{C}^n$ such that $||x|| = 1$, and $W = \{w^1, w^2\}$ i.i.d. complex Gaussians, let

$$
E^{x_0}(x) = \mathbb{E}(\langle y, |F \text{Diag}(w^1)x|^2 - |F \text{Diag}(w^2)x|^2 \rangle),
$$

where $y = \text{sign}(|F \text{Diag}(w^1)x_0|^2 - |F \text{Diag}(w^2)x_0|^2) \in \{-1, 1\}^n$. Moreover, let

$$
\hat{E}^{x_0}(x) = \frac{1}{r} \sum_{i=1}^{r} \langle y_i, |F \text{Diag}(w_i^1)x|^2 - |F \text{Diag}(w_i^2)x|^2 \rangle,
$$

$y_i = Q_{W_i}(x_0)$ and $W_i = \{(w_i^1, w_i^2)\}, i = 1 \ldots r$ are i.i.d. complex Gaussians.

In the following definition the phase retrieval problem is cast as an empirical risk maximization:

**Definition 25** (Phase retrieval Problem). The phase retrieval problem amounts to solving:

$$
\max_{x, \|x\| = 1} F_{x_0}(x)
$$

Let $\hat{x}_r = \arg \max_{x, \|x\| = 1} \hat{E}^{x_0}(x)$.

The following proposition shows that the objective function can be written explicitly as a quadratic form.

**Proposition 6.** $E^{x_0}(x)$ can be rewritten as the following quadratic form:

$$
E^{x_0}(x) = x^* C x, \quad (10.19)
$$

where $C = \mathbb{E}(\text{Diag}(w^1)F \text{Diag}(y)F^* \text{Diag}(w^{1,*}) - \text{Diag}(w^2)F \text{Diag}(y)F^* \text{Diag}(w^{2,*}))$ and

$$
\hat{E}^{x_0}(x) = x^* \hat{C}_r x, \quad (10.20)
$$

where $\hat{C}_r = \frac{1}{r} \sum_{i=1}^{r} \left(\text{Diag}(w_i^1)F \text{Diag}(y_i)F^* \text{Diag}(w_i^{1,*}) - \text{Diag}(w_i^2)F \text{Diag}(y_i)F^* \text{Diag}(w_i^{2,*})\right)$.

The phase retrieval problem from One bit CDP is therefore a maximum eigenvalue problem, that we call $1bitPhase$:

$$
\max_{x, \|x\| = 1} x^* \hat{C}_r x \quad (10.21)
$$
10.3.3 Theoretical analysis: Correctness in Expectation and Concentration

In this section we sketch the main steps of the proof of Theorem 12. The reader is referred to Section 10.7 for detailed proofs.

The following proposition shows that $x_0$ is indeed the leading eigen-vector of the expected problem (10.19) with eigen-value $\lambda$, where $\lambda$ is given in (9.28). Moreover the expected matrix $C$ is rank one:

**Proposition 7 (Correctness in Expectation).** The following statements hold:

1. For all $x \in \mathbb{C}^n$, $||x|| = 1$, we have the following equality,

$$E_{x_0}(x) = x^*Cx = \lambda|\langle x_0, x \rangle|^2.$$  \hfill (10.22)

2. Let $y = Q^g_A(x_0)$, $C$ is a rank one matrix,

$$C = \lambda x_0x_0^*.$$ \hfill (10.23)

3. $x_0$ is an eigenvector of $C$ with eigenvalue $\lambda$,

$$Cx_0 = \lambda x_0.$$ \hfill (10.24)

4. The maximum eigenvector of $C$ is of the form $x_0e^{i\phi}$, where $\phi \in [0, 2\pi]$. The maximum eigenvalue is given by $\lambda$.

The following lemma is a comparison equality that allows us to bound $||x^*x - x_0x_0^*||_F^2$, for any point $x$, by the excess risk $E_{x_0}(x_0) - E_{x_0}(x)$:

**Lemma 15.** The following equality holds for all $x \in \mathbb{C}^n$:

$$E_{x_0}(x_0) - E_{x_0}(x) = \frac{\lambda}{2}||xx^* - x_0x_0^*||_F^2.$$  \hfill (10.25)

The rest of the proof follows from empirical processes theory [LT91] and concentration inequalities [Tro12].

**Proposition 8 (Concentration).** Let $\hat{x}_r = \arg \max_{x, ||x||=1} x^*\hat{C}_r x$, the following inequalities hold:

1. \[ \frac{\lambda}{2}||\hat{x}_r^*\hat{x}_r - x_0x_0^*||_F^2 \leq 2||\hat{C}_r - C||. \]
2. For $0 < \epsilon < 1, r \geq c\frac{\log^3 n}{\epsilon^2}$, $\|\hat{C}_r - C\| \leq \epsilon$ with probability at least $1 - O(n^{-2})$.

Proof of Theorem 12. The proof of Theorem 12 follows from a simple combination of Proposition 7, Lemma 15 and Proposition 8.

Proofs of Corollaries 3 and 4 are simple consequences of Theorem 12 and Lemma 16, where we specify the value of $\lambda$ for each model.

Lemma 16. The values of $\lambda$ for different observation models $\theta$ are given in the following:

1. Noiseless setup: $\theta(z) = z$, $\lambda = 1$.

2. Noisy setup: $\theta(z) = z + v$, $v$ is an exponential random variable with variance $\sigma$, $\lambda = \frac{1 + 2\sqrt{\sigma}}{(1 + \sqrt{\sigma})^2}$.

3. Distortion setup: $\theta(z) = \tanh(\alpha z)$, where $\alpha > 0$,
   $\lambda = \mathbb{E}(\text{sign}(1 - \tanh(\alpha E_1) \tanh(\alpha E_2)) |E_1 - E_2|)$ is a decreasing function in $\alpha$.

10.4 From One bit Coded Diffraction Patterns to Super-Resolution

We turn now to the problem of recovering a signal from its lower end of power spectra. As discussed earlier this is a problem of practical interest, as the resolution of an optical system, for instance a lens $h$ is limited by the Fraunhofer diffraction limit $f_c$. The super-resolution factor of $h$ is therefore defined as $\text{SRF} = \frac{n}{2f_c + 1}$. In our setup the modulated signal diffracts through a lens characterized by a PSF $h$ and a cut-off frequency $f_c$. Hence instead of observing the power spectra of the modulated signal $\text{Diag}(w) x_0$ we observe the power spectra of a lower resolution signal namely $h \ast (\text{Diag}(w) x_0)$. Let $\hat{u}$ be the Fourier transform of $u \in \mathbb{C}^n$, $\hat{u} = Fu$. Note that by the properties of the Fourier transform we have:

$$F(h \ast (\text{Diag}(w) x_0)) = \hat{h} \otimes \text{Diag}(w) x_0$$

Hence we observe:

$$|F(h \ast (\text{Diag}(w) x_0))|^2 = |\hat{h}|^2 \otimes |\text{Diag}(w) x_0|^2$$

In this section we re-index $k$ for convenience $-\frac{n}{2} \leq k \leq \frac{n}{2} - 1$. We use also the following convention $\text{sign}(0) = 1$, this choice is arbitrary. Note that due to the diffraction limit $f_c$, $h$ satisfies:

$$|\hat{h}|^2_k = 0, \text{ for } k \notin [-f_c, f_c].$$
Hence our phase-less measurement are missing in high frequencies ranges:

\[ b_k = \hat{h}_k^2 \widehat{Diag}(w) x_{0_k}^2 \quad k \in [-f_c, f_c] \text{ and } b_k = 0 \text{ elsewhere.} \tag{10.27} \]

The one bit coded diffraction patterns are the defined by comparing pairs of low resolution coded diffraction patterns \((b^1, b^2)\) defined as in (10.27) for two independent modulations \(w^1, w^2 \sim \mathcal{CN}(0, I_n)\).

\[ y_k = \text{sign} (b_k^1 - b_k^2), \quad k = 1 \ldots n, \]

Note that \(y_k = 0\) if \(k \notin [-f_c, f_c]\). For all \(k \in [-f_c, f_c]\), we have \((d_k^1 = |FDiag(w^1)x_{0_k}^2, d_k^2 = |FDiag(w^2)x_{0_k}^2) \sim (\frac{1}{n}E_k^1, \frac{1}{n}E_k^2), \) where \(E_k^1, E_k^2\) iid \(Exp(1)\). We are now ready to compute the value of \(\lambda\) corresponding to that observation model:

\[
\lambda = \mathbb{E}(\langle y, |FDiag(w^1)x_{0}^2 - |FDiag(w^2)x_{0}^2 \rangle)
\]

\[
= \sum_{k=\frac{-n-1}{2}}^{\frac{n-1}{2}} \mathbb{E}(\text{sign}(b_k^1 - b_k^2)(d_k^1 - d_k^2)).
\]

Note that:

For \(k \in [-f_c, f_c]\) \(\text{sign}(b_k^1 - b_k^2) = \text{sign} (|\hat{h}_k^1d_k^1 - |\hat{h}_k^2d_k^2) = \text{sign} (|\hat{h}_k^2(d_k^1 - d_k^2)) = \text{sign}(d_k^1 - d_k^2)\).

\[ \tag{10.28} \]

it follows that:

\[
\lambda = \sum_{k \in [-f_c, f_c]} \mathbb{E}(\text{sign}(d_k^1 - d_k^2)(d_k^1 - d_k^2))
\]

\[ = \frac{1}{n} \sum_{k \in [-f_c, f_c]} \mathbb{E}(\text{sign}(E_k^1 - E_k^2)(E_k^1 - E_k^2))
\]

\[ = \frac{2f_c + 1}{n} \mathbb{E}( |E^1 - E^2|)
\]

\[ = \frac{2f_c + 1}{n}.
\]

\[ = \frac{1}{\text{SRF}}.
\]

Thus we find that \(\lambda\) is the inverse of the super-resolution factor, which confirms our findings on \(\lambda\) as a quality factor: \(\lambda\) is small when the cut off frequency is small. Interestingly \(\lambda\) depends only on the domain of \(\hat{h}\), regardless the shape or the values of the corresponding PSF, and this is due to the quantization step in (10.28). This the main reason behind the feasibility of blind deconvolution within our framework. Corollary 5 follows simply from Theorem 12 setting \(\lambda = \frac{1}{\text{SRF}}\).
10.5 Algorithms and Computations

10.5.1 One Bit Phase Retrieval Algorithm

A straightforward computation of the maximum eigenvector of \( \hat{C}_r \) is rather expensive. Using the power method, and the Fast Fourier transform we get a computational complexity of

\[ O(n \log^4(n)) \]

Note by FFT the Fast Fourier Transform and iFFT the Inverse Fast Fourier Transform. For the power method we need to compute \( \hat{C}_r u \):

\[
\hat{C}_r u = \frac{1}{r} \sum_{i=1}^{r} \text{Diag}(w_i^1) F \text{Diag}(y_i) F^* \text{Diag}(w_i^{1,*}) u - \text{Diag}(w_i^2) F \text{Diag}(y_i) F^* \text{Diag}(w_i^{2,*}) u
\]

\[
= \frac{1}{r} \sum_{i=1}^{r} w_i^1 \circ \text{FFT} (y_i \circ \text{iFFT}(w_i^{1,*} \circ u)) - w_i^2 \circ \text{FFT} (y_i \circ \text{iFFT}(w_i^{2,*} \circ u))
\]

For each iteration we need to compute the FFT for each pair of modulations, that costs \( O(n \log(n)) \) per pair. We have \( O(\log^3 n) \) pairs, hence a total of \( O(n \log^4(n)) \) operations per iteration of the power method.

\[
\text{Algorithm 5 } \text{FastFourier1bitCDPPhasePower}
\]

1: procedure FastFourier1bitCDPPhasePower \( \{w_i^1, w_i^2\}_{i=1...r}, y = (y_1 \ldots y_r), \epsilon \) 
2: Initialize \( r_0 \) at random, \( j = 1 \).
3: while \( ||u_j - u_{j-1}|| > \epsilon \) or \( j = 1 \) do
4: \( u_j \leftarrow \frac{1}{r} \sum_{i=1}^{r} w_i^1 \circ \text{FFT} (y_i \circ \text{iFFT}(w_i^{1,*} \circ u_{j-1})) \)
5: \( u_j^2 \leftarrow \frac{1}{r} \sum_{i=1}^{r} w_i^2 \circ \text{FFT} (y_i \circ \text{iFFT}(w_i^{2,*} \circ u_{j-1})) \)
6: \( u_j \leftarrow u_j^1 - u_j^2 \)
7: \( \lambda \leftarrow ||u_j|| \)
8: \( r_j \leftarrow \frac{u_j}{\lambda} \)
9: \( j \leftarrow j + 1 \)
10: end while
11: return \( (\lambda, u) \) \( \triangleright (\lambda, u) \) is an estimate of \( (\lambda, x_0) \).
12: end procedure

10.5.2 SubExp Initialization

In this spirit of the initialization SubExpPhase proposed in [NJS13] for Gaussian measurements we propose the following initialization from coded diffraction patterns. Let
\[ b_i = |FDiag(w_i)x_0|^2, \ i = 1 \ldots L, \ \text{where} \ w_i \sim \mathcal{CN}(0, I_n) \ \text{iid.} \]

Define

\[ \hat{C}_L = \frac{1}{L} \sum_{i=1}^{L} \text{Diag}(w_i) F \text{Diag}(b_i) F^* \text{Diag}(w_i^*), \]

It is possible to show that:

\[ \mathbb{E}(\hat{C}_L) = x_0^* x_0 + I_n, \]

we omit the proof and refer the reader to Lemma 3.1 in [CLM13] for a similar argument. As in the case of the Gaussian measurements the sample complexity of \textit{SubExpPhase} is higher than One bit Phase Retrieval. An inspection of the proof of one bit phase retrieval shows that \( O(\log^5 n/\epsilon^2) \) modulations are needed for an \( \epsilon \) recovery in \textit{SubExpPhase}(we omit the details and show that this is indeed encountered in practice in Section 10.6.1). Let \( \hat{x}_L \) be the maximum eigenvector of \( \hat{C}_L \), \( \hat{x}_L \) is a good proxy of \( x_0 \), hence the following algorithm that proceeds by the power method in order to find the maximum eigenvector:

**Algorithm 6** FastFourierSubExpCDPPhasePower

1. procedure \textsc{FastFourierSubExpCDPPhasePower}(\( \{w_i^1, w_i^2\}_{i=1}^{r}, b = ((b_1^1, b_1^2) \ldots (b_r^2, b_r^2)), \epsilon \))
2. Initialize \( r_0 \) at random, \( j = 1 \).
3. while \( ||u_j - u_{j-1}|| > \epsilon \) or \( j = 1 \) do
4. \( u_j \leftarrow \frac{1}{r} \sum_{i=1}^{r} w_i^1 \odot \text{FFT}(b_i \odot \text{iFFT}(w_i^1 \cdotp u_{j-1})) + w_i^2 \odot \text{FFT}(b_i^2 \odot \text{iFFT}(w_i^2 \cdotp u_{j-1}))) \)
5. \( \lambda \leftarrow ||u_j|| \)
6. \( r_j \leftarrow \frac{u_j}{\lambda} \)
7. \( j \leftarrow j + 1 \)
8. end while
9. return \( (\hat{\lambda}, u) \) \( \triangleright (\hat{\lambda}, u) \) is an estimate of \( (\lambda, x_0) \).
10. end procedure

**10.5.3 Alternating Minimization Initialized with the One Bit Solution or the solution of SubExpPhase**

Given the one bit solution or the solution of \textit{SubExpPhase} we can refine the solution, by running the alternating minimization procedure on the actual coded diffraction patterns initialized with the one bit solution or with the solution of \textit{SubExpPhase} as follows:
Algorithm 7 FastAltMin+OneBitCDP initialization

1: procedure ALTMINPHASE\( (w^1_1, w^2_2, \ldots, w^1_r, w^2_r, b = (\sqrt{b^1_1}, \sqrt{b^2_1}, \ldots, \sqrt{b^1_r}, \sqrt{b^2_r}), \epsilon) \)
2: \( \text{Initialize } x \leftarrow \text{FASTFOURIER1BITCDPPHASEPOWER}(\{w^1_i, w^2_i\}_{i=1}^r, y = (y_1, \ldots, y_r), \epsilon) \) or \( x \leftarrow \text{FASTFOURIERSUBEXPCDPPHASEPOWER}(\{w^1_i, w^2_i\}_{i=1}^r, b = (b_1, \ldots, b_2r), \epsilon) \)
3: for \( k = 1 \ldots t_0 \) do \( \triangleright t_0 \) is the number of iterations.
4: \( (u^1_1, u^2_1, \ldots, u^1_r, u^2_r) \leftarrow (\text{Ph}(\text{FFT}(w^1_1 \odot x)), \text{Ph}(\text{FFT}(w^2_1 \odot x)), \ldots) \)
5: \( x \leftarrow \sum_{i=1}^r \frac{1}{|w^1_i|^2 + |w^2_i|^2} \odot \left( \sum_{i=1}^r w^1_i \ast \text{iFFT}(\sqrt{b^1_i} \odot u^1_i) + w^2_i \ast \text{iFFT}(\sqrt{b^2_i} \odot u^2_i) \right) \)
6: end for
7: return \( x \)
8: end procedure

The computational complexity of Algorithm 7 is \( O(n \log^4 n) \) by iteration. The full analysis of this algorithm is subject to future research.

10.6 Numerical Experiments

10.6.1 One Dimensional Simulations

In this section we test our algorithms on one dimensional simulated signals. We consider \( x_0 \in \mathbb{R}^n \), such that \( x_0 \) is a gaussian vector, \( x_0 \sim \mathcal{N}(0, I_n) \), we set \( n = 8000 \). We first study the phase transition of One Bit Phase Retrieval and compare it to its counterpart in \( \text{SubExpPhase} \) in Section 10.6.1. We then show in Section 10.6.1 the robustness of One Bit phase retrieval, to noise, distortion and blur.

Phase Transition of One Bit Phase Retrieval Versus SubExpPhase

We consider in this section how the performance of Algorithm 5 for \( \text{1bitPhase} \) and Algorithm 6 for \( \text{SubExpPhase} \) depend on the number of measurements. We consider 50 trials, where we generate pairs of CDP \( (b^1_{\ell}, b^2_{\ell}), \ell = 1 \ldots r \) according to Equation (10.11) and the one bit CDP \( y_{\ell}, \ell = 1 \ldots r \), according to (10.12), where we set \( \theta \) to be the identity (noiseless model). At each trial we generate a set of new random modulation, and run Algorithms 5 (\( \text{1bitPhase} \)) and 6 (\( \text{SubExpPhase} \)).

In Figure 10-1, we report the empirical probability of success of each algorithm for increasing number of measurements \( r \). We have a success if the error \( 1 - |\langle \hat{x}_r, x_0 \rangle|^2 < \tau \). We set \( n = 8000, \tau = 0.07 \) in this experiment. We see that the phase transition for \( \text{1bitPhase} \) happens earlier than the one for \( \text{SubExpPhase} \), which confirms that One Bit Phase retrieval allows lower sample complexity for a given precision.
Figure 10-1: Phase transition comparison of one bit Phase retrieval and SubExp phase Retrieval.

**Robustness**

We test the robustness of Algorithm 5 to noise, distortion, and blur. For this end we generate measurements according to the noise model given in Equation (10.7), for increasing noise level and for $r = 10$ and $r = 20$. We see in Figure 10-2(a) that that the recovery error of One Bit Phase Retrieval increases gracefully with the level of noise and as more measurements are available the error of recovery drops down. In the setup of distorted measurements we generate CDP according to Equation (10.10), for different distortion levels $\alpha$, for different number of measurements $r = 10$, and $r = 20$. We see in Figure 10-2(b) that the recovery is still possible thanks to the robustness of One Bit Phase Retrieval, despite the severe non linearity. Finally we test the robustness of 1bitPhase to a gaussian blur with increase aperture. We generate our CDP according to the blurry model in Equation (10.10), for $r = 10$ and $r = 20$. We see in Figure 10-2(c) that phase retrieval ad Super-Resolution are possible with One Bit Phase Retrieval and that the error increases gracefully also with the size of the aperture and drops as more measurements are available.
I

\[ -r = 20 \]

\[ 0.2 - r = 20 0.2 \]

\[ 0 0.5 1 1.5 2 0 20 40 60 80 100 \]

\[ 0 \]

\[ g \]

\[ 0 \]

\[ (a) \text{ Robustness to Exponential} \]

\[ (b) \text{ Robustness to distortion: } Er - \]

\[ (c) \text{ Robustness to blur: } \text{Error} 1 - \]

\[ \text{noise: } \text{Error} 1 - |\langle x, x_0 \rangle|^2 \text{ versus } \mu \text{ or } r - |\langle x, x_0 \rangle|^2 \text{ versus } \sigma \text{ the aperture of the mean of the Exponential noise of the clipping}. \]

\[ \text{a gaussian filter.} \]

Figure 10-2: Phase transition comparison of one bit Phase retrieval and SubExp phase Retrieval.

One Bit Phase Retrieval and Alternating Minimization

Alternating Minimization Phase Transition. As discussed in Section 10.2.1, and Section 10.5.3, greedy refinements of the solution of 1bitPhase and SubExpPhase, enhance the quality of the recovery and the sample complexity of the overall procedure as showed in [NJS13] for the Gaussian measurements. Extending those results to Coded Diffraction patterns is subject to future work. We conjecture on one hand that AM initialized with the one bit solution has a sample complexity of \( O(\log^3 n + \log \frac{1}{\epsilon} \log \log \frac{1}{\epsilon}) \), and on the other hand that AM initialized with SubExpPhase solution has a sample complexity of \( O(\log^5 n + \log \frac{1}{\epsilon} \log \log \frac{1}{\epsilon}) \). We show in the next section the phase transition of AM (Algorithm 7) initialized with 1bitPhase, SubExpPhase and a random initialization in the noisy and the noiseless case.

In order to highlight the effect of the initialization step in the AM Algorithm 7, we fix the number of iterations to \( t_0 = 50 \) in the noiseless and the noisy setting.

In figure 10-3(a) we report the empirical success probability of Algorithm 7 in the noiseless setting versus the number of pairs of modulations \( r \). We declare in the noiseless case a success if \( ||\hat{x}_{t_0}^{*} - x_0x_0^{*}||_F < 10^{-5} \). We see that with that relatively small number of iterations, the phase transition of AM initialized with One Bit Phase solution happens at \( r = 4 \). AM initialized with SubExpPhase with that limited number of iterations needs more samples to achieve phase transition at \( r = 10 \). AM initialized at random does not achieve its phase transition with that limited number of samples and iterations. This phase transition confirms the lower sample complexity of one bit solution, and its greedy refinements.

We now turn to the noisy setting (10.7) where we set \( \sigma = 0.04 \). In figure 10-3(b) we report the empirical success probability of Algorithm 7, where we declare in this setting a success if \( ||\hat{x}_{t_0}^{*} - x_0x_0^{*}||_F < 0.03 \). We see that at that accuracy level, the greedy refinements of one bit solutions are still robust to noise and superior to the other forth-mentioned initializations (SubExpPhase and random).

155
Figure 10-3: Phase transition comparison of one bit Phase retrieval and SubExp phase retrieval.

**Error Decay.** To illustrate the benefit of the initialization step in AM we report in the following the error decay of AM with different initializations. In figure 10-4(a) we see that in the noiseless setting all approaches converge, the convergence is faster for AM initialized with the one bit solution in high dimension. In figure 10-4(b), 10-4(c), 10-4(d) we see that AM initialized with one bit solution is more robust in the noisy setting.
Figure 10-4: Alternating minimization convergence with different initializations: Random Initialization, 1bitPhase, and SubExpPhase, in the noisy and noiseless setting.

In the next section we test our algorithms in imaging applications, which highlights the efficiency and the robustness of the algorithms in potential applications in microscopy, astronomy and X-ray Crystallography.

10.6.2 Imaging Applications

We address in this section the problem of phase recovery in imaging applications. In the following we consider two test images, and their respective power spectra. The image in
Figure 10-5 has a dominant edge structure, and the image in Figure 10-6 has a dominant textured content. We show that the same algorithms presented in this chapter allow the phase recovery, where we simply replace the vectors with 2D arrays and the 1D FFT with the 2D FFT.

Figure 10-5: An image of Stata Center (Structured) and its Power Spectra (we plot the logarithm of the power spectra of one color Channel (R for instance)).

Figure 10-6: A drawing of Stata Center (texture like image) and its Power Spectra (we plot the logarithm of the power spectra of one color Channel (R for instance)).

One Bit Coded Diffraction Patterns

We modulate the image with a 2D Gaussian array and collect 2D coded diffraction patterns. We illustrate in Figure 10-7 the one bit coded diffraction patterns that become in this case a 2D binary array. The one bit array is obtained by quantizing pairs of coded diffraction arrays. Our goal is therefore to recover robustly the image from the knowledge of One bit
coded diffraction arrays using the same algorithms presented in this chapter. In the next section we test the robustness of the recovery against distortion, noise and blur.

Figure 10-7: One Bit Coded Diffraction Patterns obtained by quantizing pairs of coded diffractions patterns.

**Phase Retrieval from Distorted Diffraction Patterns.**

In many acquisition systems we collect the power spectra of an object of interest. This acquisition might be altered by many imperfections due to multiple scattering phenomena for instance, or distortion in the precision of the CCD. Robustness to distortion such as clipping is a desirable feature in phase retrieval. In Figure 10-8 we simulate distorted power spectra by applying a sigmoid ($\tanh(\alpha)$) to the Fourier spectrum, for different clipping levels ($\alpha$). While most of the spatial frequency information is lost, phase retrieval is still possible thanks to the robustness of One Bit Phase retrieval to distortions. We apply $2r$ Gaussian masks to our image of interest and then collect the power spectra of each masked image. We apply to each masked power spectra a sigmoid with a clipping parameter $\alpha$. We the get our One Bit CDP by quantizing pairs of distorted power spectras. In our experiment $n = 256 \times 256$, $r = 16[\log(n)]$. 

159
Figure 10-8: Stata Center's power spectra undergoing a distortion such as clipping of the values \( \text{tanh}(\alpha|\hat{x}(w)|) \). The distorted power spectra is obtained for \( \alpha = 0.001 \). The severely distorted is obtained for \( \alpha = 0.1 \).

Figure 10-9: One Bit Phase Retrieval from distorted CDP. (a) and (b) Recovery from small distortions. (c) Recovery from Severe Distortions.

We run independently the same process of acquisition on the three color channels, as well as Algorithm 5. In Figures 10-9 and 10-10 we show the output of the Algorithm 5 for various level of distortions varying from mild to severe distortion. We see that the reconstruction in both cases is still possible from one bit CDP despite the distortion that the intensities values are undergoing.
Super-Resolution and Blind Deconvolution

We turn now to the problem of recovering an image from its lower end of power spectra. As discussed earlier this is a problem of practical interest in microscopy, as the resolution of an optical system, for instance a lens $h$, is limited by the Fraunhofer diffraction limit $f_c$. The super-resolution factor of $h$ is therefore defined as $SRF = \frac{n}{2f_c+1}$. In our setup the modulated object diffracts through a lens characterized by a PSF $h$ and a cut-off frequency $f_c$. Hence instead of observing the power spectra of the modulated signal $\text{Diag}(w)x_0$ we observe the power spectra of a lower resolution signal namely $h \ast (\text{Diag}(w)x_0)$. We consider in this experiment $h$ to be an averaging filter. In Figure 10-11(a) we see a blurred image, obtained by convolving the original image Stata with an averaging filter of size $8 \times 8$. In the following we simulate the diffraction patterns of the modulated image through the aperture $h$ by taking: $|Fh \ast (\text{Diag}(w_i)x_0)|^2, |Fh \ast (\text{Diag}(w_i)x_0)|^2, i = 1 \ldots r$. We then obtain the one bit CDP by quantizing pairs of CDPs. In our experiment we have $n = 256 \times 256, r = 10 \lfloor \log(n) \rfloor$.

In Figure 10-11(b), we show the output of Algorithm 5 given the one bit CDP collected as mentioned previously. We see that most of the missing details in Figure 10-11(a) are recovered. Hence one bit phase retrieval enables super-resolution and blind deconvolution as it is agnostic to the nature of the blur.
One Bit Phase Retrieval and Alternating Minimization

In this section we test the alternating minimization Algorithm 7 initialized with One Bit Phase Retrieval (Algorithm 5), in the noiseless and the Poisson noise model of equation (10.8). To emphasize the effect of the initialization we set the number of iterations $t_0$ in Algorithm 7 to a relatively small number. In our experiments we set $t_0 = 50$, and $r = 4$. We test our algorithms with Complex Gaussian masks and Bernoulli masks.

**Gaussian masks.** We start by the recovery for a noiseless acquisition of the CDP, the recovered image with alternating minimization initialized with one bit phase retrieval is given in Figure 10-12(a) and is indistinguishable from the original. The average SNR on the three color channels is 101.2248 dB. When the CDP are contaminated with a poisson noise (In Equation (10.8) we set $\eta = 0.1$), Alternating minimization initialized with one bit phase retrieval succeeds and produces a solution with an average SNR on the three color channels of 95.625 dB. The recovered image in this setting is shown Figure 10-12(b).

**Bernoulli masks.** We repeat the same experiment with bernoulli masks, i.e each entry of the mask is a bernoulli random variable with parameter $p = 0.8$. The Recovered image in the noiseless is given in Figure 10-13(a), the average SNR is 97 dB. In the poisson noise setting the recovered image is given in Figure 10-13(b), the average SNR is 95.0769 dB. We see that the quality for this fixed number of masks and iterations is lower than the Gaussian case.
(a) Recovered Image with AM initialized with the 1bit-Phase solution, for a noiseless acquisition of CDP. (b) Recovered Image with AM initialized with the 1bit-Phase solution, for an acquisition where the CDP are contaminated with a Poisson Noise.

Figure 10-12: Alternating Minimization and One Bit Phase Retrieval from Gaussian masks.

(a) Recovered Image with AM initialized with the 1bit-Phase solution, for a noiseless acquisition of CDP. (b) Recovered Image with AM initialized with the 1bit-Phase solution, for an acquisition where the CDP are contaminated with a Poisson Noise.

Figure 10-13: Alternating Minimization and One Bit Phase Retrieval from Bernoulli masks.

10.7 Proofs

In this section we give the proofs of Proposition 7, Lemma 15 and Proposition 8. We start with the following observation:
Lemma 17. Let $y \in \mathbb{C}^n$, and $M \in \mathbb{C}^{n \times n}$, we have the following equality:

$$\langle y, \text{diag}(M) \rangle = \text{Tr}(\text{diag}(y)^*M).$$

(10.29)

We need this result for $y \in \mathbb{R}^n$:

$$\langle y, \text{diag}(M) \rangle = \text{Tr}(\text{diag}(y)M).$$

(10.30)

Let $v = F^*\text{Diag}(w^*)x$ and $u = F^*\text{Diag}(w^*)x_0$. The proof of proposition 6 follows from Lemmas 18 and 19:

Lemma 18. The modulus vector can be rewritten in the following way:

$$|F\text{Diag}(w)x|^2 = \text{diag}\{F^*\text{Diag}(w^*)xx^*\text{Diag}(w)F\} = \text{diag}(vv^*).$$

Lemma 19. The following equation holds:

$$\langle y, |F\text{Diag}(w)x|^2 \rangle = x^*\text{Diag}(w)F\text{Diag}(y)F^*\text{Diag}(w^*)x.$$

Proof of Proposition 6. By Lemma 19

$$\mathcal{E}^{x_0}(x) = \mathbb{E}(\langle y, |F\text{Diag}(w^1)x|^2 - |F\text{Diag}(w^2)x|^2 \rangle)$$

$$= x^*Cx.$$

where $C = \mathbb{E}(\text{Diag}(w^1)F\text{Diag}(y)F^*\text{Diag}(w^{1,*}) - \text{Diag}(w^2)F\text{Diag}(y)F^*\text{Diag}(w^{2,*})).$

Proof of Lemma 18.

$$|F\text{Diag}(w)x|^2[i] = \sum_j F_{ij}w_jx_j \sum_k \bar{F}_{ik}\bar{w}_k\bar{x}_k = \sum_{jk} F_{ij}\bar{F}_{ik}\bar{w}_k w_j x_j \bar{x}_k.$$ 

(10.31)

On the other hand:

$$e_i^*F^*\text{Diag}(w^*)xx^*\text{Diag}(w)Fe_i = [\bar{F}_{i1} \ldots \bar{F}_{in}]W[F_{i1} \ldots F_{in}]',$$

(10.32)

where $W_{jk} = w_j\bar{w}_k x_j \bar{x}_k$. Hence:

$$e_i^*F^*\text{Diag}(w^*)xx^*\text{Diag}(w)Fe_i = \sum_{jk} F_{ij}\bar{F}_{ik}\bar{w}_k w_j x_j \bar{x}_k.$$ 

(10.33)

The lemma is proved.
Proof of Lemma 19. Using Lemma 17 and Lemma 18, we have:

\[ \langle y, |F\text{Diag}(w)x|^2 \rangle = \langle y, \text{diag}\{F^*\text{Diag}(w^*)xx^*\text{Diag}(w)F\} \rangle \]
\[ = \text{Tr}(\text{Diag}(y)F^*\text{Diag}(w^*)xx^*\text{Diag}(w)F) \]
\[ = \text{Tr}(\text{Diag}(w)F\text{Diag}(y)F^*\text{Diag}(w^*)xx^*) \]
\[ = x^*\text{Diag}(w)F\text{Diag}(y)F^*\text{Diag}(w^*)x. \]

\[ \square \]

Proof of Proposition 7. i-Let \( v = F^*\text{Diag}(w^*)x \) and \( u = F^*\text{Diag}(w^*)x_0, u_i \) and \( u_i \sim \mathcal{CN}(0, \frac{1}{n}), i = 1 \ldots n. \) \( u \) and \( v \) are Gaussian vectors with dependent coordinates. The expectation of dot product \( \langle u, v \rangle \) is given in the following:

\[ \mathbb{E}(\langle u, v \rangle) = \mathbb{E}(x_0^*\text{Diag}(w)FF^*\text{Diag}(w^*)x) = x_0^*\mathbb{E}(\text{Diag}(|w|^2))x = \langle x_0, x \rangle. \]

since \( FF^* = I, \) and \( \mathbb{E}(\text{Diag}(|w|^2)) = I_n. \)
Hence \( \mathbb{E}(|u|^2) = \mathbb{E}(|v|^2) = 1, \) since \( x \) and \( x_0 \) are unitary.
There exists a complex Gaussian random vector \( r, \) such that each \( r_i \sim \mathcal{CN}(0, \frac{1}{n}) \) and \( r_i \) is independent of \( u_i, i = 1 \ldots n \) and:

\[ v = \langle x_0, x \rangle u + \sqrt{1 - \langle x, x_0 \rangle^2}r. \]

Let \( v_1 = F^*\text{Diag}(w^1*)x, u_1 = F^*\text{Diag}(w^1*)x_0, \) and \( v_2 = F^*\text{Diag}(w^2*)x, u_2 = F^*\text{Diag}(w^2*)x_0. \)
By lemma 18 we have:

\[ \mathcal{E}^{x_0}(x) = \mathbb{E}(\langle y, |F\text{Diag}(w^1)x|^2 - |F\text{Diag}(w^2)x|^2 \rangle) \]
\[ = \mathbb{E}(\langle y, \text{diag}(v_1v_1^*-v_2v_2^*) \rangle) \]

On the other hand:

\[ vv^* = (\langle x_0, x \rangle u + \sqrt{1 - \langle x, x_0 \rangle^2}r)(\langle x_0, x \rangle u^* + \sqrt{1 - \langle x, x_0 \rangle^2}r^*) \]
\[ = |\langle x_0, x \rangle|^2uu^* + (1 - |\langle x, x_0 \rangle|^2)rr^* + 2\sqrt{1 - |\langle x_0, x \rangle|^2}\text{Re}(\langle x_0, x \rangle ur^*) \]

Therefore:

\[ \text{diag}(v_1v_1^*-v_2v_2^*) = |\langle x_0, x \rangle|^2\text{diag}(u_1u_1^*-u_2u_2^*) + (1 - |\langle x, x_0 \rangle|^2)\text{diag}(r_1r_1^*-r_2r_2^*) \]
\[ + 2\sqrt{1 - |\langle x, x_0 \rangle|^2}\text{diag}(\text{Re}(\langle x_0, x \rangle(u_1r_1^*-u_2r_2^*))) \]
\[ = |\langle x_0, x \rangle|^2(|F\text{Diag}(w)x_0|^2 - |F\text{Diag}(w^2)x_0|^2) \]
\[ + (1 - |\langle x, x_0 \rangle|^2)\text{diag}(r_1r_1^*-r_2r_2^*) \]
\[ + 2\sqrt{1 - |\langle x, x_0 \rangle|^2}\text{diag}(\text{Re}(\langle x_0, x \rangle(u_1r_1^*-u_2r_2^*))). \]
Therefore, taking the expectation we have:

$$\mathbb{E}^{x_0}(x) = |\langle x, x_0 \rangle|^2 \mathbb{E} \left( \langle y, |F\text{Diag}(w^1)x_0| - |F\text{Diag}(w^2)x_0| \rangle \right)$$

since $\mathbb{E}_{u_1,u_2} \langle y, \mathbb{E}_{r_1,r_2}(\text{diag}(r_1r_1^* - r_2r_2^*)|u_1, u_2) \rangle = 0,$
and $\mathbb{E}_{u_1,u_2} (\langle y, \mathbb{E}_{r_1,r_2}(\text{diag}(\text{Re}(u_1r_1^* - ur_2^*))|u_1, u_2) \rangle) = 0.$

Let $
\lambda = \mathbb{E} \left( \langle y, |F\text{Diag}(w^1)x_0| - |F\text{Diag}(w^2)x_0| \rangle \right).
$

Let $\tilde{E}_i^1 = |\sum_{j=1}^n F_{ij} w_j x_0, j|^2,$ since $w_j$ is complex Gaussian, $\sum_{j=1}^n F_{ij} w_j x_0, j$ is $\mathcal{CN}(0, \frac{1}{n}),$
therefore $\tilde{E}_i^1$ is exponential with mean $\frac{1}{n}$ (since $|F_{ij}|^2 = 1$). Let $E_i^1$ be an exponential with
mean 1. $\tilde{E}_i^1 = \frac{1}{n} E_i^1.$ It follows that:

$$\lambda = \mathbb{E} \left( \langle y, |F\text{Diag}(w^1)x_0| - |F\text{Diag}(w^2)x_0| \rangle \right)$$

$$= \frac{1}{n} \sum_{i=1}^n \mathbb{E}(E_i^1 - E_i^1)$$

$$= \frac{1}{n} n \mathbb{E}(E^1 - E^2)$$

$$= \mathbb{E}(E^1 - E^2)$$

Where $\lambda$ is defined for different models in Lemma 16.

**Remark 7.** Similar results hold if we use instead of the DFT $F$ any unitary matrix $U.$ The
value of $\lambda$ would depends on $|U_{ij}|^2.$

**Proof of Proposition 8.** It follows that:

$$\mathbb{E}(\tilde{C}_r) = \lambda x_0 x_0^*$$

Where $\tilde{C}_r = \frac{1}{r} \sum_{i=1}^r A_i$
where $A_i = \text{Diag}(w_1^i) F\text{Diag}(y_i) F^* \text{Diag}(w_1^{i,*}) - \text{Diag}(w_2^i) F\text{Diag}(y_i) F^* \text{Diag}(w_2^{i,*})$  $i = 1 \ldots r.$

By Lemma 15, it is now clear that the sample complexity is governed by the concentration
of $\tilde{C}_r$ around its mean. Let $E_\beta = \{ w \in \mathbb{C}^n, |w_1|^2 \leq 2 \beta \log(n) \text{ and } |w_2|^2 \leq 2 \beta \log(n), j = 1 \ldots n \}.$ Let $(w_1^1, w_2^1), i = 1 \ldots r,$ be $2r$ independent iid $\mathcal{CN}(0, I_n).$ Define $(\tilde{w}_1^1, \tilde{w}_2^1) = (w_1^1, w_2^1)$
if $(w_1^1, w_2^1) \in E_\beta$ and $(\tilde{w}_1^1, \tilde{w}_2^1) = (0, 0)$ elsewhere.

Define $\tilde{C}_r = \frac{1}{r} \sum_{i=1}^r \tilde{A}_i,$ where

$$\tilde{A}_i = \text{Diag}(\tilde{w}_1^i) F\text{Diag}(\tilde{y}_i) F^* \text{Diag}(\tilde{w}_1^{i,*}) - \text{Diag}(\tilde{w}_2^i) F\text{Diag}(\tilde{y}_i) F^* \text{Diag}(\tilde{w}_2^{i,*}),$$

and let $\tilde{C} = \mathbb{E}\tilde{C}_r.$ By the triangular inequality we have:

$$||\tilde{C}_r - C|| \leq ||\tilde{C}_r - \tilde{C}_r|| + ||\tilde{C}_r - \tilde{C}|| + ||\tilde{C} - C||$$ (10.34)
Bounding $\|\hat{C}_r - \hat{C}_r\|$:

Note that $\|\hat{C}_r - \hat{C}_r\| = 0$, if for all $i = 1 \ldots r$, $(w_i^1, w_i^2) \in E_\beta$. Let us get a bound on the probability of that event. Note that: $P(|w_i|^2 > 2\beta \log(n)) \leq 2n^{-\beta}$. To avoid cumbersome notations when we use index $i$, $w_i$ refers to a modulation in $\mathbb{C}^n$, $i = 1 \ldots r$, and when we use index $j$ $w_j$ refers to the $j$-th component of $w \in \mathbb{C}^n, j = 1 \ldots n$.

\[
P \{ \exists i \in \{1 \ldots r\} \text{ such that } (w_i^1, w_i^2) \notin E_\beta \} \leq r P \{ (w^1, w^2) \notin E_\beta \} = r P \{ \bigcup_{t=1,2} \{ \exists j \text{ such that } |w_j^t|^2 > 2\beta \log(n) \} \} \leq \frac{4rn}{n^\beta}.
\]

It follows that:

$$\|\hat{C}_r - \hat{C}_r\| = 0 \text{ with probability at least } 1 - \frac{4r}{n^{\beta - 1}}. \quad (10.35)$$

Bounding $\|\hat{C}_r - \hat{C}\|$:

Let

$$\tilde{X}_i = \tilde{A}_i - \mathbb{E}(\tilde{A}_i),$$

$\mathbb{E}(\tilde{X}_i) = 0$. Note that

$$\|\hat{C}_r - \hat{C}\| = \frac{1}{r} \left\| \sum_{i=1}^r \tilde{X}_i \right\|$$

Let us get a bound on $\|\tilde{X}_i\|$. Note that $\|\mathbb{E}(\tilde{A}_i)\| \leq \lambda$. To simplify the notation we will omit in the following the indices.

$$\left\| \tilde{A} \right\| = \sup_{x, \|x\|=1} \langle \tilde{y}, |F \text{Diag}(\tilde{w})^1 x|^2 - |F \text{Diag}(\tilde{w})^2 x|^2 \rangle$$

Recall that $v = F^* \text{Diag}(\tilde{w}^*) x$. Note that: $|F \text{Diag}(\tilde{w}) x|^2 = \text{diag} \{ F^* \text{Diag}(\tilde{w})^1 xx^* \text{Diag}(\tilde{w}) F \} = \text{diag}(vv^*)$.

By holder inequality we have:

$$\langle \tilde{y}, |F \text{Diag}(\tilde{w})^1 x|^2 - |F \text{Diag}(\tilde{w})^2 x|^2 \rangle = \langle \tilde{y}, \text{diag}(v^1 v^1) - \text{diag}(v^2 v^2) \rangle \leq \left\| \tilde{y} \right\|_\infty \left\| \text{diag}(v^1 v^1) - \text{diag}(v^2 v^2) \right\|_{\ell_1}.$$

Since $\tilde{y}$ is binary $\|\tilde{y}\|_\infty = 1$. By the triangular inequality:

$$\left\| \text{diag}(v^1 v^1) - \text{diag}(v^2 v^2) \right\|_{\ell_1} \leq \left\| \text{diag}(v^1 v^1) \right\|_{\ell_1} + \left\| \text{diag}(v^2 v^2) \right\|_{\ell_1}.$$
\[ \| \text{diag}(vv^*) \|_{\ell_1} = \| \text{diag} \{ F^* \text{diag}(\tilde{w}^*)xx^* \text{diag}(\tilde{w})F \} \|_{\ell_1} = \text{Tr}(F^* \text{diag}(\tilde{w}^*)xx^* \text{diag}(\tilde{w})F) = \text{Tr}(\text{diag}(\tilde{w})FF^* \text{diag}(\tilde{w}^*)xx^*) = \text{Tr}(\text{diag}(|\tilde{w}|^2)xx^*) \leq \| \text{diag}(|\tilde{w}|^2) \| \| x \|^2. \]

We are now left with Bounding:

\[
\sup_{x, ||x||=1} \left( \| \text{diag}(|\tilde{w}^1|^2) \| + \| \text{diag}(|\tilde{w}^2|^2) \| \right) \| x \|^2 = \left( \| \text{diag}(|\tilde{w}^1|^2) \| + \| \text{diag}(|\tilde{w}^2|^2) \| \right) = \max_{j=1\ldots n} |\tilde{w}^1_j|^2 + \max_{j=1\ldots n} |\tilde{w}^2_j|^2 \quad (10.36)
\]

By definition of \((\tilde{w}^1, \tilde{w}^2)\) we conclude that:

\[ ||\tilde{A}|| \leq 4\beta \log(n). \quad (10.37) \]

It follows that

\[ ||\tilde{X}_i|| \leq 4\beta \log(n) + \lambda \leq 5\beta \log(n) := \Delta. \]

**Theorem 13** (Hoeffding Matrix Inequality [Tro12]). Let \( X_i, i = 1\ldots r \) be a sequence of independent random \( n \times n \) self-adjoint matrices. Assume that each random matrix obeys:

\[ \mathbb{E}(X_i) = 0 \quad \text{and} \quad ||X_i|| \leq \Delta \text{ almost surely}. \]

Then for all \( t \geq 0 \),

\[ \mathbb{P} \left\{ \frac{1}{r} \left\| \sum_{i=1}^r X_i \right\| \geq t \right\} \leq 2n \exp \left( -\frac{rt^2}{8\Delta^2} \right). \]

In other words:

For \( r \geq \frac{c^2t^2}{\epsilon^2 \Delta^2 \log(n)} \), \( \frac{1}{r} \left\| \sum_{i=1}^r X_i \right\| \leq \epsilon \) with probability at least \( 1 - n^{-c^2} \).

We are now ready to apply the Hoeffding Matrix inequality:

For \( r \geq c t^2 \beta^2 \log^3 n \), \( \frac{1}{r} \left\| \sum_{i=1}^r \tilde{X}_i \right\| \leq \epsilon \) with probability at least \( 1 - n^{-c^2} \).

It follows that:

\[ \| \tilde{C}_r - \tilde{C} \| \leq ct\beta \sqrt{\frac{\log^3 n}{r}} \text{ with probability at least } 1 - n^{-c^2}. \quad (10.38) \]

168
Bounding $||\hat{C} - C||$:

By Jensen inequality followed by Cauchy Schwarz inequality we have:

\[
||\hat{C} - C|| = ||\mathbb{E}(\mathbb{I}_{w_1, w_2 \notin E_\beta} A)|| \\
\leq \mathbb{E}\mathbb{I}_{w_1, w_2 \notin E_\beta} ||A|| \\
\leq \sqrt{\mathbb{E}(\mathbb{I}_{w_1, w_2 \notin E_\beta})} \sqrt{\mathbb{E}(||A||^2)} \\
\leq \sqrt{\mathbb{P}(E_\beta^c)} \sqrt{\mathbb{E}(\max_{j=1...n} |w_j|^2 + \max_{j=1...n} |w_j|^2)^{2}}.
\]

The last inequality follows from equation (10.36).

\[
||A||^2 \leq (\max_{j=1...n} |w_j|^2)^2 + (\max_{j=1...n} |w_j|^2)^2 + 2 \max_{j=1...n} |w_j|^2 \max_{j=1...n} |w_j|^2.
\]

(10.39)

Let $Z = \max_{j=1...n} E_j$, $E_j \sim \text{Exp}(1)$ iid, therefore:

\[
\mathbb{E}(||A||^2) \leq 2(E(Z)^2 + (E(Z))^2) = 2(\text{Var}(Z)) + 2(E(Z))^2).
\]

(10.40)

**Lemma 20** (Maximum of Exponential [BT12]). Let $Z = \max_{j=1...n} E_j$, $E_j \sim \text{Exp}(1)$ iid, therefore: $\text{Var}(Z) \leq 2 \mathbb{E}(Z) = \sum_{i=1}^{n} \frac{1}{i} \leq \log(n)$.

For sufficiently large $n$, there exists a constant $c$ such that:

\[
\mathbb{E}(||A||^2) \leq 2(E(Z)^2 + (E(Z))^2) = 2(2 + 2(\log^2(n))) = 4(1 + \log^2(n)) \leq c^2 \log^2(n).
\]

(10.41)

Note that $\mathbb{P}(E_\beta^c) \leq \frac{4n}{n^3}$. Therefore:

\[
\left|\left|\hat{C} - C\right|\right| \leq \frac{2c}{n^{(\beta-1)/2}} \log(n).
\]

(10.42)

**Putting all together:**

Putting together equations (10.34),(10.35),(10.38) and (10.42) we have finally with probability at least $1 - n^{-t^2} - \frac{4r}{n^{\beta-1}}$:

\[
\left|\left|\hat{C}_r - C\right|\right| \leq c\beta t \sqrt{\frac{\log^3 n}{r}} + \frac{2c}{n^{(\beta-1)/2}} \log(n).
\]

(10.43)

Setting $t = \sqrt{2}, \beta = 4$ we get with probability $1 - O(n^{-2})$,

\[
\left|\left|\hat{C}_r - C\right|\right| \leq 4\sqrt{2} c \sqrt{\frac{\log^3 n}{r}} + \frac{2c}{n^{3/2}} \log(n).
\]

(10.44)
It follows that there exists a numeric constant $c$ such that:

\[ \text{For } r \geq c \frac{\log^3 n}{\epsilon^2} \quad \| \hat{C}_r - C \| \leq \epsilon, \text{ with probability at least } 1 - O(n^{-2}). \quad (10.45) \]

Finally by equation (7.24) we conclude that:

\[ \text{For } r \geq c \frac{\log^3 n}{\epsilon^2} \quad \frac{1}{2} \| \hat{x}_r \hat{x}_r^* - x_0x_0^* \|_F^2 \leq \frac{\epsilon}{\lambda}, \text{ with probability at least } 1 - O(n^{-2}). \quad (10.46) \]

In other words for another numeric constant $c$ we have:

\[ \text{For } r \geq c \frac{\log^3 n}{\epsilon^2 \lambda^2} \quad \| \hat{x}_r \hat{x}_r^* - x_0x_0^* \|_F^2 \leq \epsilon, \text{ with probability at least } 1 - O(n^{-2}). \quad (10.47) \]

**Lemma 21.** For $x, x_0 \in \mathbb{S}^{n-1}$, assume $\langle x, x_0 \rangle = \cos(\theta), \theta \in \left[0, \frac{\pi}{2}\right]$. Let $0 < \epsilon < 1$, and $\phi \in [0, 2\pi]$ we have:

\[ \frac{1}{2} \| xx^* - x_0x_0^* \|_F^2 \leq \epsilon \implies \frac{1}{2} \| x - x_0e^{i\phi} \|_F^2 \leq \epsilon. \]

**Proof.** $\theta \in \left[0, \frac{\pi}{2}\right], 0 < \epsilon < 1, \text{ we have:}$

\[ \frac{1}{2} \| xx^* - x_0x_0^* \|_F^2 = 1 - \cos^2(\theta) \leq \epsilon \]

Since $\theta \in \left[0, \frac{\pi}{2}\right], \cos^2(\theta) \geq 1 - \epsilon \implies 1 - \cos(\theta) \leq 1 - \sqrt{1 - \epsilon}.$

For $0 < x < 1$, we have $\sqrt{x} \geq x$, hence we conclude that:

\[ \frac{1}{2} \| x - x_0e^{i\phi} \|_F^2 = 1 - \cos(\theta) \leq 1 - (1 - \epsilon) = \epsilon. \]

$\Box$
Chapter 11

Conclusions and Perspectives

In this thesis we show that following a coding/decoding approach, we can learn efficiently semantic mapping of the data in the context of multiclass and multimodal classification. Following the same quantization approach we show that a compressive representations of the data, based on vector quantization of non linear random measurements $\theta((a, x))$, can be defined on a $q$-ary alphabet. This representation can be used in the context of locality sensitive hashing in information retrieval, its unquantized counterpart can be used for local linear approximation of function in a regression or classification task. This representation is invertible for structured signals and a decoder can be defined using $q-$ ary compressive sensing, a generalization of the growing field of one bit compressive sensing. An important feature of this compressive representation goes to its robustness to unknown perturbation $\theta$. Finally we define a binary quantization scheme for phase-less linear measurement under-going a possible non linearity $\theta, \theta(|\langle a, x \rangle|)$, and show that a redundant representation, based on those binary measurements, allow robust phase retrieval, and super-resolution beyond the diffraction limit in the context of the Fourier based sampling. Quantization offers in that case a good initial point for greedy algorithms, and play the role of a preconditioning that enhances the sample complexity of the overall optimization. We believe that quantization would help in defining good initial points to other non convex problems in signal processing and machine learning.

We wish to highlight some perspectives and open questions related to the framework built in this thesis:

1. Improved rates for one bit phase retrieval from one bit CDP using Bernstein inequalities akin to [GKK14] and the intrinsic dimension. We believe that optimal rate is $O(\log(n))$.

2. Recovering the norm in one bit and $q$- ary quantization in compressive sensing and phase retrieval using random thresholds or the distribution of the quantized measurements.

3. Improved rates for phase-lift form CDP, by constructing a dual certificate that uses quantized measurements in the first step in the golfing scheme [GKK14].
4. Proof of convergence of alternating minimization for coded diffraction pattern and the one bit initialization.

5. One Bit coded diffraction patterns and non Gaussian modulations or masks akin to one bit compressive sensing from non Gaussian measurements \([ALPV12]\) using the Lindeberg swapping trick.

6. An open question in the interaction between quantized compressive sensing and Learning theory, remains on how to learn a provable binary embedding from data \([RFD09, RG07, BD09, WTF08]\)?

7. Most deep learning architectures \([ALR^{+}13, LBBH98, BM12, VLB08, HSK^{+}12]\) consist in the succession of redundant or compressive embeddings, followed by averaging a non linearity \(\theta\) applied to those measurements. The sampling vectors can be learned from data in an unsupervised or supervised way. Random sampling vectors have been also used successfully in that context \([ALR^{+}13, LKF10]\). More formally a deep or hierarchical embedding \(C\) consists in the composition of many non linear embeddings \(C_h, h = 1 \ldots H, C = C_H \circ \cdots \circ C_1\), such that:

\[
C_h: \mathbb{R}^{n_h} \rightarrow \mathbb{R}^{|G_\ell|} \rightarrow \mathbb{R}^{L_h}
\]

\[
u \mapsto \{\langle a_j, u \rangle, j \in G_\ell\}_{\ell=1 \ldots L_h} \mapsto C_h(u) = \left\{ \sum_{j \in G_\ell} \theta(\langle a_j, x \rangle) \right\}_{\ell=1 \ldots L_h}
\]

or in the case of a complex sensing set \([BM12]\):

\[
u \mapsto \{\langle a_j, u \rangle, j \in G_\ell\}_{\ell=1 \ldots L_h} \mapsto C_h(u) = \left\{ \sum_{j \in G_\ell} \theta(|\langle a_j, x \rangle|) \right\}_{\ell=1 \ldots L_h}
\]

where \(n_1 = n\), and \(n_{h+1} = L_h\).

In an unsupervised setting an open question is what is the condition on the randomness, the groups \(G_\ell\), the signal class \(K\) such that there exists \(D_1 \ldots D_H\) satisfying:

**Reconstruction requirement:**

\[
\text{Let } v = C_H \circ \cdots \circ C_1(x_0), \quad \sup_{x_0 \in K} \|D_1 \circ \cdots \circ D_H(v) - x_0\|_2 < \epsilon,
\]

this requirement is the basis of sparse coding based, Independent Component Analysis, Independent Subspace Analysis or auto-encoders deep networks \([VLBM08]\).
- Contraction requirement:

$$\mathbb{E}_{g} \sup_{x_0, x_1 \in K} \langle g, C_H \circ \cdots \circ C_1(x_0) - C_H \circ \cdots \circ C_1(x_1) \rangle \leq (1 - e^\alpha) \mathbb{E}_{g'} \sup_{x_0, x_1 \in K} \langle g', x_0 - x_1 \rangle,$$

where $\alpha < 0$, $g' \sim \mathcal{N}(0, I_n)$, and $g \sim \mathcal{N}(0, I_{LH})$.

The intuition behind the requirement on the contraction in the Gaussian complexity, is that the host space should be simpler than the original one, while there exists an inverse mapping from the host space to the original space, as a guarantee that there was no loss in information. Are those requirement enough to guarantee in the unsupervised case a good embedding? Similar questions can be asked if the signals were drawn from a probability distribution $\rho$ instead of belonging to a signal set $K$, replacing the sup with an expectation.

In the supervised setting we can ask the question of what is a good feature map? i.e what is a good discriminative embedding? Let $V$ one of the multi-class losses introduced, we shall minimize:

$$\min_{f, C_1, \ldots, C_H} \mathbb{E} \left( V \left( y, f(C_H \circ \cdots \circ C_1(x)) \right) \right),$$

this is a challenging problem to study and optimize, an unsupervised initialization is usually used in deep neural (convolutional) networks [LKF10, HSK+12]. An open question is on the consistency of this procedure and its convergence.
Bibliography


