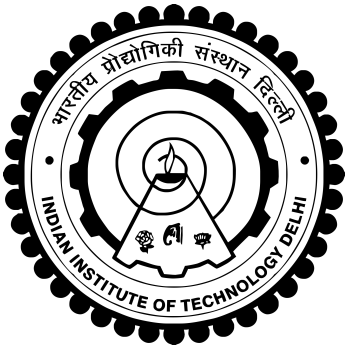
LARGE SCALE AND SPARSE MINIMAL COMPLEXITY MACHINES

MAYANK SHARMA



DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY DELHI
OCTOBER 2019

LARGE SCALE AND SPARSE MINIMAL COMPLEXITY MACHINES

by

MAYANK SHARMA

Department of Electrical Engineering

Submitted

in fulfilment of requirements for the degree of Doctor of Philosophy

to the

INDIAN INSTITUTE OF TECHNOLOGY DELHI
OCTOBER 2019

Certificate

This is to certify that the thesis entitled "**Large scale and sparse minimal complexity machines**", being submitted by **Mayank Sharma** for the award of the degree of **Doctor of Philosophy** to the Department of Electrical Engineering, Indian Institute of Technology Delhi, is a record of bonafide work done by him under my supervision and guidance. The matter embodied in this thesis has not been submitted to any other University or Institute for the award of any other degree or diploma.

Dr. Jayadeva
Professor
Department of Electrical Engineering,
Indian Institute of Technology Delhi,
Hauz Khas, New Delhi - 110016,
INDIA.

Acknowledgments

I would like to thank my supervisor Prof. Jayadeva, Research Committee members Professors Santanu Chaudhary, Amit Kumar and Sumeet Agarwal; Professor Suresh Chandra for his guidance; friends and colleagues Sumit Soman, Himanshu Pant, Prashant Gupta, Aashi Jindal; department staff members Rakesh, Yatindra, Mukesh and Ritwick; my parents, brother and especially my fiancée Ms. Divya Vatsa for supporting me throughout this journey.

(*Mayank Sharma*)

Abstract

The capacity of a learning machine is characterized by various measures such as the Vapnik-Chervonenkis (VC) dimension, Rademacher complexity, and covering numbers. All these measures define a notion of *complexity* or capacity of a learning machine in either how well it can fit a random labelling of data points or how many samples it can classify with all possible labellings of the data points. Several generalization bounds are proposed in the literature using these measures. Large capacity can lead to overfitting, while a small one can lead to under-fitting. Researchers have worked on developing several algorithms and paradigms to find the right balance of empirical error and capacity. One such work in this direction is the Minimal Complexity Machine (MCM). The MCM was formulated as a way to tightly bound the VC dimension. The MCM was shown to outperform the Support Vector Machine (SVM) in generating sparse solutions and resulting in excellent generalization properties. In this thesis, we extend the MCM and develop several algorithms to scale up the MCM to large datasets while maintaining sparsity, lower complexity, and having a small generalization error. In doing so, we identified four major problems with the state-of-the-art SVM and the least squares SVM (LS-SVM). First, they have a massive space complexity. Second, they do not generate highly sparse solutions. Third, they cannot use *indefinite kernels* and last they are not suitable for IoT devices due to large model size even after quantization.

We develop solutions to the problems mentioned above around the recently proposed MCM. We initially develop a Stochastic Gradient Descent (SGD) variant of the MCM that can scale to large datasets via the use of *explicit feature maps*. However, the use of *explicit feature maps* does not provide the information about the points that make up the decision boundary and hence, we develop algorithms in the Empirical Feature Space (EFS). The EFS is defined to be all linear combinations of vectors in the kernel or succinctly, the *span* of the kernel. We define the MCM in the EFS, the least squares MCM and their respective input space margin maximization variants. The advantage of using EFS is that the number of training samples bounds the VC dimension of a classifier in the EFS. We use *prototype vector* selection of data points in scaling the EFS variants to large datasets.

The EFS variants developed use One-vs-One (OVO) classification scheme for multiclass classification scheme, which results in slower training and inference time for datasets with a large number of classes. To tackle this issue, we develop novel multiclass MCM variants along with their least squares and margin maximization variants in the EFS. We show that these variants offer the advantage of faster training and prediction time while maintaining sparsity and accuracies higher than the OVO variants. We also extend the multiclass MCM to a GPU based acceleration hardware, where we scale the MCM to > 1 million data points.

The next part of the thesis focuses on extending the multiclass MCM variants

to using *indefinite kernels*. The EFS formulations allow for the use of indefinite kernels in the optimization problem of the multiclass MCM. We transition from the Hilbert spaces to Banach and Krein spaces to present analysis for the indefinite kernel multiclass MCM and the least squares MCM. We demonstrate that indefinite kernel MCM variants have similar generalization properties as the positive definite kernel MCM variants.

In the last part of the thesis, we solve the final issue identified by us in extending the algorithms to memory constraint IoT devices. We present weight quantization comparison of the multiclass MCM and the SVM variants such as the least squares SVM and the LIBSVM. We evaluate all our algorithms on 13 benchmark datasets, spanning small and large datasets both in terms of the number of samples and classes. We show that the proposed multiclass MCM variants can retain accuracies even using 3 bits as compared to 12 bits of used by the SVM and the LS-SVM.

सार

एक मशीन के सीखने की क्षमता की तुलना विभिन्न उपायों जैसे कि वैपनिक-चेरोवेनेकिस (वीसी) आयाम, रेडेमैकर जटिलता, और कवर संख्याओं के आधार पे की जाती है । इन सभी उपायों में एक मशीन की जटिलता या क्षमता को उसकी यादृच्छिक लेबलिंग को सीखने की क्षमता या सभी उदाहरणों के सभी संभावित लेबलिंग के वर्गीकरण करने की दक्षता के आधार पे परिभाषित किया जाता है। इन उपायों का उपयोग करके साहित्य में कई सामान्यीकरण सीमाएँ प्रस्तावित हैं। बड़ी क्षमता से ओवरफिटिंग हो सकती है, जबकि एक छोटे से अंडर-फिटिंग हो सकती है। शोधकर्ताओं ने अनुभवजन्य त्रुटि और क्षमता का सही संतुलन खोजने के लिए कई एल्गोरिदम और प्रतिमान विकसित करने पर काम किया है। इस दिशा में ऐसा ही एक काम है मिनिमल कॉम्प्लेक्सिटी मशीन (एमसीएम)। एमसीएम को वीसी आयाम को कसने के लिए एक तरीके के रूप में तैयार किया गया था।

एमसीएम को विरल समाधान पैदा करने में सपोर्ट वेक्टर मशीन (एसवीएम) से बेहतर दिखाया गया था और जिसके परिणामस्वरूप एमसीएम में उत्कृष्ट सामान्यीकरण गुण थे। इस थीसिस में, हम एमसीएम का विस्तार बड़े डेटासेट के लिए करते हुए इसे विरल, कम जटिल, और सामान्यीकरण त्रुटि को सामान्य रखने का प्रयास करते है और इस दिशा में कई एल्गोरिथ्म का अनुसंधान करते हैं । ऐसा करते हुए, हमने अत्याधुनिक एसवीएम और लीस्ट स्क्वायर एसवीएम (एलएस-एसवीएम) के साथ चार प्रमुख समस्याओं की पहचान की। सबसे पहले, उनकी स्पेस जटिलता काफी अधिक है। दूसरा, वे अत्यधिक विरल समाधान उत्पन्न नहीं करते हैं। तीसरा, वे अनिश्चित कर्नेल का उपयोग नहीं कर सकते हैं और अन्ततः वे क्वान्टिजेशन के बाद भी बड़े आकार के उत्प्पन मॉडल के कारण (आई ओ टी) उपकरणों के लिए उपयुक्त नहीं हैं।

हम हाल ही में प्रस्तावित एमसीएम के आसपास उपर्युक्त समस्याओं के समाधान विकसित करते हैं। हम शुरू में एम सीए एम का एक स्टोकेस्टिक ग्रेडिएंट डिसेंट (एस जी डी) वेरिएंट विकसित करते हैं जो बड़े डेटासेट को स्पष्ट फीचर मैप्स के उपयोग के माध्यम से स्केल कर सकता है। हालांकि, स्पष्ट फीचर मैप्स का उपयोग उन उदाहरणों के बारे में जानकारी प्रदान नहीं करता है जो निर्णय सीमा बनाते हैं और इसलिए, हम एम्पिरिकल फीचर स्पेस (ई एफ एस) में एल्गोरिदम विकसित करते हैं। ई एफ एसको कर्नेल के सभी रैखिक संयोजनों के रूप में परिभाषित किया गया है या संक्षेप में, कर्नेल का स्पैन है। हम ई एफ एस में एम सीए एम , लीस्ट स्क्वायर एम सीए एम और उनके संबंधित इनपुट स्पेस मार्जिन को बढ़ने वाले विकल्प को परिभाषित करते हैं। ईएफएस का उपयोग करने का लाभ यह है कि ईएफएस में वर्गीकारक का वीसी आयाम प्रशिक्षण उदाहरणों की संख्या से घिरा हुआ है। हम ई एफ एस वेरिएंट को बड़े डेटासेट में स्केल करने में उदाहरणों के प्रोटोटाइप वेक्टर चयन का उपयोग करते हैं।

ई एफ एस के मल्टीक्लास वेरिएंट वर्गीकरण योजना के लिए वन-बनाम-वन (ओ वी ओ) वर्गीकरण योजना का उपयोग करते हैं , जिसके परिणामस्वरूप प्रषिखण और अनुमान का समय अत्यधिक हो जाता है । इस समस्या के समाधान के लिए, हम ई एफ एस का नवीन मल्टीक्लास एमसीएम, लीस्ट स्क्वायर तथा मार्जिन अधिकतमकरण के वैरिएंट विकसित करते हैं। हम दिखाते हैं कि ये वेरिएंट ओ वी ओ वेरिएंट की तुलना में अधिकता और सटीकता को बनाए रखते हुए तेज़ प्रशिक्षण और अनुमान के समय का लाभ देते हैं। हम मल्टीकॉलस एमसीएम को जीपीयू आधारित त्वरण हार्डवेयर में भी विस्तारित करते हैं, जहां हम एमसीएम को> 1 मिलियन डेटा उदाहरणों पर स्केल करते हैं।

थीसिस का अगला भाग अनिश्चित कर्नेल का उपयोग करते हुए मल्टीक्लास एम सीए एम वेरिएंट को विस्तारित करने पर केंद्रित है। ई एफ एस योगों को मल्टीक्लास एम सीए एमकी अनुकूलन समस्या में अनिश्चित कर्नेल के उपयोग की अनुमति देता है। हम हिल्बर्ट स्पेस से बानच और क्रेन स्पेस में अनिश्चितकालीन कर्नेल मल्टीस्केल्स एमसीएम और लीस्ट स्क्वायर एमसीएम के लिए विश्लेषण प्रस्तुत करते हैं। हम प्रदर्शित करते हैं कि अनिश्चित कर्नेल एमसीएम वेरिएंट में सकारात्मक निश्चित कर्नेल एमसीएम वेरिएंट के समान सामान्यीकरण गुण होते हैं ।

थीसिस के अंतिम भाग में, हम एल्गोरिदम को मेमोरी बाधिक आईओटी उपकरणों तक पहुंचाने में हमारे द्वारा पहचाने गए अंतिम मुद्दे को हल करते हैं। हम मल्टीक्लास एमसीएम और एसवीएम वेरिएंट जैसे लीस्ट स्क्वायर एसवीएम और एलआईबीएसवीएम के पैरामीटर क्वान्टिजेशन की तुलना करते हैं। हम 13 बेंचमार्क डेटासेट पर अपने सभी एल्गोरिदम का मूल्यांकन उदाहरणों तथा वर्गों की संख्या के सन्दर्भ में करते हैं । हम दिखाते हैं कि प्रस्तावित मल्टीक्लास एमसीएम वेरिएंट एसवीएम और एलएस-एसवीएम द्वारा उपयोग किए जाने वाले 12 बिट्स की तुलना में 3 बिट्स का उपयोग करके भी सटीकता बनाए रखता है

Contents

Acknowledgments **iii**

Abstract **v**

List of Figures **xiii**

List of Tables **xviii**

List of Equations **xx**

1 Introduction **1**
1 Background . . . 1
1.1 Goals of the thesis . . . 2
2 Measure of Complexity: Vapnik-Chervonenkis Dimension . . . 3
3 Support Vector Machines . . . 5
3.1 Rademacher Complexity Bounds . . . 8
4 Minimal Complexity Machines . . . 9
5 Organization of the thesis . . . 10
5.1 Contributions . . . 10
5.2 Thesis structure . . . 11

2 Large Scale Minimal Complexity Machines Based On Explicit Feature Maps **13**
1 Introduction . . . 13
2 The Minimal Complexity Machine . . . 15
2.1 Kernel Minimal Complexity Machine . . . 17
2.2 Kernel L_1 norm Minimal Complexity Machine . . . 18
3 The Stochastic Sub-Gradient descent MCM . . . 19
3.1 Random Fourier Features . . . 20
3.2 Nystrom . . . 21
3.3 Fastfood . . . 21
3.4 Tensor Sketches . . . 21
3.5 Stochastic Sub-Gradient descent Non-linear MCM . . . 22
3.6 Convergence . . . 22
4 Results . . . 25
4.1 Feature normalization and Hyperparameter tuning . . . 26
4.2 Results on small datasets . . . 27
4.3 Results on large datasets . . . 29
4.4 p-values comparison . . . 29
5 Conclusion and Future Work . . . 32

3 Ultra Sparse Minimal Complexity Machines Based On Empirical Feature Maps 33
1 Introduction . . . 33
2 The Empirical Feature Space . . . 34
2.1 The EFS-MCM . . . 36
2.2 The EFS-MCM-M . . . 40
2.3 The EFS-LS-MCM . . . 42
2.4 The EFS-LS-MCM-M . . . 43
2.5 Large scale EFS-MCM and EFS-LS-MCM . . . 44
3 Results . . . 53
3.1 Feature normalization and Hyperparameter tuning . . . 53
3.2 Results on small datasets for non least squares versions . . . 54
3.3 Results on large datasets for non least squares versions . . . 56
3.4 p-values comparison for non least squares versions . . . 56
3.5 Results on small datasets for least squares versions . . . 58
3.6 Results on large datasets for least squares versions . . . 60
3.7 p-values comparison for least squares versions . . . 62
4 Conclusion & Future Work . . . 63

4 Multiclass Sparse Minimal Complexity Machines Based On Empirical Feature Maps 65
1 Introduction . . . 65
2 The $\mathcal{T}$-EFS-MCM . . . 65
3 The $\mathcal{T}$-EFS-LS-MCM . . . 70
4 The $\mathcal{T}$-EFS-MCM-M . . . 71
5 The $\mathcal{T}$-EFS-LS-MCM-M . . . 71
6 Large scale variants . . . 72
7 Results . . . 72
7.1 Feature normalization and Hyperparameter tuning . . . 78
7.2 Results on small datasets for non least squares versions . . . 80
7.3 Results on large datasets for non least squares versions . . . 82
7.4 p-values comparison for non least squares versions . . . 82
7.5 Results on small datasets for least squares versions . . . 83
7.6 Results on large datasets for least squares versions . . . 85
7.7 p-values comparison for least squares versions . . . 87
7.8 GPU implementation of the $\mathcal{T}$-EFS-MCM-SGD . . . 90
8 Conclusion & Future Work . . . 92

5 Non-Mercer Minimal Complexity Machines 95
1 Introduction . . . 95
2 Existing approaches . . . 96
2.1 Spectrum modification . . . 96
2.2 Solvers accepting indefinite matrices . . . 96
3 Empirical Feature Space . . . 97
4 Multiclass formulations . . . 101
4.1 The $\mathcal{T}$-EFS-MCM and the $\mathcal{T}$-EFS-LS-MCM . . . 101
4.2 The $\mathcal{T}$-EFS-MCM-M and the $\mathcal{T}$-EFS-LS-MCM-M . . . 102
4.3 Large scale formulations . . . 103

5 Results . . . 104
5.1 Feature normalization and Hyperparameter tuning . . . 107
5.2 Results on small datasets for non least squares versions . . . 107
5.3 Results on large datasets for non least squares versions . . . 110
5.4 p-values comparison for least squares versions . . . 113
5.5 Results on small datasets for least squares versions . . . 118
5.6 Results on large datasets for least squares versions . . . 121
5.7 p-values comparison for least squares versions . . . 124
6 Conclusion . . . 127

6 Weight Quantized Minimal Complexity Machines 129
1 Introduction . . . 129
2 Parameter Quantization . . . 130
2.1 Quantization scheme for ILP EFS based algorithms . . . 130
2.2 Post training quantization scheme for EFS based algorithms . 131
3 Algorithms . . . 132
4 Bounds on functional margin for post-training quantization . . . 136
5 Results . . . 137
5.1 Feature normalization and Hyperparameter tuning . . . 138
5.2 Results on small datasets . . . 139
5.3 Results on large datasets . . . 158
6 Conclusion . . . 168

7 Conclusion 173
1 Future Works . . . 175

Acronyms 177

List of Publications 200
2 Publications related to thesis chapters . . . 200
3 Co-Authored Journal Publications . . . 200
4 Co-Authored Conference Publications . . . 200
5 Preprints . . . 201

Brief Biodata of the Author 203

List of Figures

2.1 Summary of the MCM for large datasets. 15
2.2 Box plots of mean test set accuracies (a) and mean log2 run time (s) (b) for small datasets 28
2.3 Box plots of mean test set accuracies (a) and mean log2 run time (s) (b) for small datasets 30

3.1 Box plots of mean test set accuracies (a), mean log2 number of support vectors (b), mean log2 L_2 norm (c) and mean log2 run time (s) (d) across small datasets for non least squares formulations. 55
3.2 Box plots of mean test set accuracies (a), mean log2 number of support vectors (b), mean log2 L_2 norm (c) and mean log2 run time (s) (d) across large datasets for non least squares formulations. 57
3.3 Box plots of mean test set accuracies (a), mean log2 number of support vectors (b), mean log2 L_2 norm (c) and mean log2 run time (s) (d) across small datasets for least squares formulations. 59
3.4 Box plots of mean test set accuracies (a), mean log2 number of support vectors (b), mean log2 L_2 norm (c) and mean log2 run time (s) (d) across large datasets for least squares formulations. 61

4.1 Box plots of mean test set accuracies (a), mean log2 number of support vectors (b), mean log2 L_2 norm (c) and mean log2 run time (s) (d) across small datasets for non least squares formulations. 79
4.2 Box plots of mean test set accuracies (a), mean log2 number of support vectors (b), mean log2 L_2 norm (c) and mean log2 run time (s) (d) across small datasets for non least squares formulations. 81
4.3 Box plots of mean test set accuracies (a), mean log2 number of support vectors (b), mean log2 L_2 norm (c) and mean log2 run time (s) (d) across small datasets for least squares formulations. 84
4.4 Box plots of mean test set accuracies (a), log2 number of support vectors (b), log2 L_2 norm (c) and log2 run time (s) (d) across large datasets for least squares formulations. 86

5.1 Box plots of mean test set accuracies (a), mean log2 number of support vectors (b), mean log2 L_2 norm (c) and mean log2 run time (s) (d) across small datasets for non least squares formulations. 108
5.1 Box plots of mean test set accuracies (a), mean log2 number of support vectors (b), mean log2 L_2 norm (c) and mean log2 run time (s) (d) across small datasets for non least squares formulations (cont.). . 109
5.2 Box plots of mean test set accuracies (a), mean log2 number of support vectors (b), mean log2 L_2 norm (c) and mean log2 run time (s) (d) across large datasets for non least squares formulations. 111

5.2 Box plots of mean test set accuracies (a), mean log2 number of support vectors (b), mean log2 L_2 norm (c) and mean log2 run time (s) (d) across large datasets for non least squares formulations (cont.). . 112

5.3 Box plots of mean test set accuracies (a), mean log2 number of support vectors (b), mean log2 L_2 norm (c) and mean log2 run time (s) (d) across small datasets for least squares formulations. 119

5.3 Box plots of mean test set accuracies (a), mean log2 number of support vectors (b), mean log2 L_2 norm (c) and mean log2 run time (s) (d) across small datasets for least squares formulations (cont.). 120

5.4 Box plots of mean test set accuracies (a), mean log2 number of support vectors (b), mean log2 L_2 norm (c) and mean log2 run time (s) (d) across large datasets for least squares formulations. 122

5.4 Box plots of mean test set accuracies (a), mean log2 number of support vectors (b), mean log2 L_2 norm (c) and mean log2 run time (s) (d) across large datasets for least squares formulations (cont.). 123

6.1 Heatmap of mean test set accuracies for the SVM-OVO (a), the $\mathcal{T}$-EFS-MCM-SGD (b), the $\mathcal{T}$-EFS-MCM-M-SGD (c), the SFS-LS-SVM (d) $\mathcal{T}$-EFS-LS-MCM-SGD (e), the $\mathcal{T}$-EFS-LS-MCM-M-SGD (f), the EFS-MCM-ILP (g) and the EFS-LS-MCM-ILP (h) for various quantization levels (integer and fraction bits) on the small datasets. 142

6.1 Heatmap of mean test set accuracies for the SVM-OVO (a), the $\mathcal{T}$-EFS-MCM-SGD (b), the $\mathcal{T}$-EFS-MCM-M-SGD (c), the SFS-LS-SVM (d) $\mathcal{T}$-EFS-LS-MCM-SGD (e), the $\mathcal{T}$-EFS-LS-MCM-M-SGD (f), the EFS-MCM-ILP (g) and the EFS-LS-MCM-ILP (h) for various quantization levels (integer and fraction bits) on the small datasets (cont.). 143

6.1 Heatmap of mean test set accuracies for the SVM-OVO (a), the $\mathcal{T}$-EFS-MCM-SGD (b), the $\mathcal{T}$-EFS-MCM-M-SGD (c), the SFS-LS-SVM (d) $\mathcal{T}$-EFS-LS-MCM-SGD (e), the $\mathcal{T}$-EFS-LS-MCM-M-SGD (f), the EFS-MCM-ILP (g) and the EFS-LS-MCM-ILP (h) for various quantization levels (integer and fraction bits) on the small datasets (cont.). 144

6.1 Heatmap of mean test set accuracies for the SVM-OVO (a), the $\mathcal{T}$-EFS-MCM-SGD (b), the $\mathcal{T}$-EFS-MCM-M-SGD (c), the SFS-LS-SVM (d) $\mathcal{T}$-EFS-LS-MCM-SGD (e), the $\mathcal{T}$-EFS-LS-MCM-M-SGD (f), the EFS-MCM-ILP (g) and the EFS-LS-MCM-ILP (h) for various quantization levels (integer and fraction bits) on the small datasets (cont.). 145

6.2 Heatmap of log2 mean model size in bits for the SVM-OVO (a), the $\mathcal{T}$-EFS-MCM-SGD (b), the $\mathcal{T}$-EFS-MCM-M-SGD (c), the SFS-LS-SVM (d) $\mathcal{T}$-EFS-LS-MCM-SGD (e), the $\mathcal{T}$-EFS-LS-MCM-M-SGD (f), the EFS-MCM-ILP (g) and the EFS-LS-MCM-ILP (h) for various quantization levels (integer and fraction bits) on the small datasets. . 146

6.2 Heatmap of log2 mean model size in bits for the SVM-OVO (a), the $\mathcal{T}$-EFS-MCM-SGD (b), the $\mathcal{T}$-EFS-MCM-M-SGD (c), the SFS-LS-SVM (d) $\mathcal{T}$-EFS-LS-MCM-SGD (e), the $\mathcal{T}$-EFS-LS-MCM-M-SGD (f), the EFS-MCM-ILP (g) and the EFS-LS-MCM-ILP (h) for various quantization levels (integer and fraction bits) on the small datasets (cont.). 147

6.2 Heatmap of log2 mean model size in bits for the SVM-OVO (a), the $\mathcal{T}$-EFS-MCM-SGD (b), the $\mathcal{T}$-EFS-MCM-M-SGD (c), the SFS-LS-SVM (d) $\mathcal{T}$-EFS-LS-MCM-SGD (e), the $\mathcal{T}$-EFS-LS-MCM-M-SGD (f), the EFS-MCM-ILP (g) and the EFS-LS-MCM-ILP (h) for various quantization levels (integer and fraction bits) on the small datasets (cont.). 148

6.2 Heatmap of log2 mean model size in bits for the SVM-OVO (a), the $\mathcal{T}$-EFS-MCM-SGD (b), the $\mathcal{T}$-EFS-MCM-M-SGD (c), the SFS-LS-SVM (d) $\mathcal{T}$-EFS-LS-MCM-SGD (e), the $\mathcal{T}$-EFS-LS-MCM-M-SGD (f), the EFS-MCM-ILP (g) and the EFS-LS-MCM-ILP (h) for various quantization levels (integer and fraction bits) on the small datasets (cont.). 149

6.3 Heatmap of log2 mean L_2 norm for the SVM-OVO (a), the $\mathcal{T}$-EFS-MCM-SGD (b), the $\mathcal{T}$-EFS-MCM-M-SGD (c), the SFS-LS-SVM (d) $\mathcal{T}$-EFS-LS-MCM-SGD (e), the $\mathcal{T}$-EFS-LS-MCM-M-SGD (f), the EFS-MCM-ILP (g) and the EFS-LS-MCM-ILP (h) for various quantization levels (integer and fraction bits) on the small datasets. 150

6.3 Heatmap of log2 mean L_2 norm for the SVM-OVO (a), the $\mathcal{T}$-EFS-MCM-SGD (b), the $\mathcal{T}$-EFS-MCM-M-SGD (c), the SFS-LS-SVM (d) $\mathcal{T}$-EFS-LS-MCM-SGD (e), the $\mathcal{T}$-EFS-LS-MCM-M-SGD (f), the EFS-MCM-ILP (g) and the EFS-LS-MCM-ILP (h) for various quantization levels (integer and fraction bits) on the small datasets (cont.). . . 151

6.3 Heatmap of log2 mean L_2 norm for the SVM-OVO (a), the $\mathcal{T}$-EFS-MCM-SGD (b), the $\mathcal{T}$-EFS-MCM-M-SGD (c), the SFS-LS-SVM (d) $\mathcal{T}$-EFS-LS-MCM-SGD (e), the $\mathcal{T}$-EFS-LS-MCM-M-SGD (f), the EFS-MCM-ILP (g) and the EFS-LS-MCM-ILP (h) for various quantization levels (integer and fraction bits) on the small datasets (cont.). . . 152

6.3 Heatmap of log2 mean L_2 norm for the SVM-OVO (a), the $\mathcal{T}$-EFS-MCM-SGD (b), the $\mathcal{T}$-EFS-MCM-M-SGD (c), the SFS-LS-SVM (d) $\mathcal{T}$-EFS-LS-MCM-SGD (e), the $\mathcal{T}$-EFS-LS-MCM-M-SGD (f), the EFS-MCM-ILP (g) and the EFS-LS-MCM-ILP (h) for various quantization levels (integer and fraction bits) on the small datasets (cont.). . . 153

6.4 Heatmap of log2 mean training time + quantization time for the SVM-OVO (a), the $\mathcal{T}$-EFS-MCM-SGD (b), the $\mathcal{T}$-EFS-MCM-M-SGD (c), the SFS-LS-SVM (d) $\mathcal{T}$-EFS-LS-MCM-SGD (e), the $\mathcal{T}$-EFS-LS-MCM-M-SGD (f), the EFS-MCM-ILP (g) and the EFS-LS-MCM-ILP (h) for various quantization levels (integer and fraction bits) on the small datasets. 154

6.4 Heatmap of log2 mean training time + quantization time for the SVM-OVO (a), the $\mathcal{T}$-EFS-MCM-SGD (b), the $\mathcal{T}$-EFS-MCM-M-SGD (c), the SFS-LS-SVM (d) $\mathcal{T}$-EFS-LS-MCM-SGD (e), the $\mathcal{T}$-EFS-LS-MCM-M-SGD (f), the EFS-MCM-ILP (g) and the EFS-LS-MCM-ILP (h) for various quantization levels (integer and fraction bits) on the small datasets (cont.). 155
6.4 Heatmap of log2 mean training time + quantization time for the SVM-OVO (a), the $\mathcal{T}$-EFS-MCM-SGD (b), the $\mathcal{T}$-EFS-MCM-M-SGD (c), the SFS-LS-SVM (d) $\mathcal{T}$-EFS-LS-MCM-SGD (e), the $\mathcal{T}$-EFS-LS-MCM-M-SGD (f), the EFS-MCM-ILP (g) and the EFS-LS-MCM-ILP (h) for various quantization levels (integer and fraction bits) on the small datasets (cont.). 156
6.4 Heatmap of log2 mean training time + quantization time for the SVM-OVO (a), the $\mathcal{T}$-EFS-MCM-SGD (b), the $\mathcal{T}$-EFS-MCM-M-SGD (c), the SFS-LS-SVM (d) $\mathcal{T}$-EFS-LS-MCM-SGD (e), the $\mathcal{T}$-EFS-LS-MCM-M-SGD (f), the EFS-MCM-ILP (g) and the EFS-LS-MCM-ILP (h) for various quantization levels (integer and fraction bits) on the small datasets (cont.). 157
6.5 Heatmap of mean test set accuracies for the SVM-OVO (a), the $\mathcal{T}$-EFS-MCM-SGD (b), the $\mathcal{T}$-EFS-MCM-M-SGD (c), the SFS-LS-SVM (d) $\mathcal{T}$-EFS-LS-MCM-SGD (e), the $\mathcal{T}$-EFS-LS-MCM-M-SGD (f) for various quantization levels (integer and fraction bits) on the large datasets. 159
6.5 Heatmap of mean test set accuracies for the SVM-OVO (a), the $\mathcal{T}$-EFS-MCM-SGD (b), the $\mathcal{T}$-EFS-MCM-M-SGD (c), the SFS-LS-SVM (d) $\mathcal{T}$-EFS-LS-MCM-SGD (e), the $\mathcal{T}$-EFS-LS-MCM-M-SGD (f) for various quantization levels (integer and fraction bits) on the large datasets (cont.). 160
6.5 Heatmap of mean test set accuracies for the SVM-OVO (a), the $\mathcal{T}$-EFS-MCM-SGD (b), the $\mathcal{T}$-EFS-MCM-M-SGD (c), the SFS-LS-SVM (d) $\mathcal{T}$-EFS-LS-MCM-SGD (e), the $\mathcal{T}$-EFS-LS-MCM-M-SGD (f) for various quantization levels (integer and fraction bits) on the large datasets (cont.). 161
6.6 Heatmap of log2 mean model size in bits for the SVM-OVO (a), the $\mathcal{T}$-EFS-MCM-SGD (b), the $\mathcal{T}$-EFS-MCM-M-SGD (c), the SFS-LS-SVM (d) $\mathcal{T}$-EFS-LS-MCM-SGD (e), the $\mathcal{T}$-EFS-LS-MCM-M-SGD (f) for various quantization levels (integer and fraction bits) on the large datasets. 162
6.6 Heatmap of log2 mean model size in bits for the SVM-OVO (a), the $\mathcal{T}$-EFS-MCM-SGD (b), the $\mathcal{T}$-EFS-MCM-M-SGD (c), the SFS-LS-SVM (d) $\mathcal{T}$-EFS-LS-MCM-SGD (e), the $\mathcal{T}$-EFS-LS-MCM-M-SGD (f) for various quantization levels (integer and fraction bits) on the large datasets (cont.). 163
6.6 Heatmap of log2 mean model size in bits for the SVM-OVO (a), the $\mathcal{T}$-EFS-MCM-SGD (b), the $\mathcal{T}$-EFS-MCM-M-SGD (c), the SFS-LS-SVM (d) $\mathcal{T}$-EFS-LS-MCM-SGD (e), the $\mathcal{T}$-EFS-LS-MCM-M-SGD (f) for various quantization levels (integer and fraction bits) on the large datasets (cont.). 164

6.7 Heatmap of log2 mean L_2 norm for the SVM-OVO (a), the $\mathcal{T}$-EFS-MCM-SGD (b), the $\mathcal{T}$-EFS-MCM-M-SGD (c), the SFS-LS-SVM (d) $\mathcal{T}$-EFS-LS-MCM-SGD (e), the $\mathcal{T}$-EFS-LS-MCM-M-SGD (f) for various quantization levels (integer and fraction bits) on the large datasets. 165
6.7 Heatmap of log2 mean L_2 norm for the SVM-OVO (a), the $\mathcal{T}$-EFS-MCM-SGD (b), the $\mathcal{T}$-EFS-MCM-M-SGD (c), the SFS-LS-SVM (d) $\mathcal{T}$-EFS-LS-MCM-SGD (e), the $\mathcal{T}$-EFS-LS-MCM-M-SGD (f) for various quantization levels (integer and fraction bits) on the large datasets (cont.). 166
6.7 Heatmap of log2 mean L_2 norm for the SVM-OVO (a), the $\mathcal{T}$-EFS-MCM-SGD (b), the $\mathcal{T}$-EFS-MCM-M-SGD (c), the SFS-LS-SVM (d) $\mathcal{T}$-EFS-LS-MCM-SGD (e), the $\mathcal{T}$-EFS-LS-MCM-M-SGD (f) for various quantization levels (integer and fraction bits) on the large datasets (cont.). 167
6.8 Heatmap of log2 mean training time + quantization time in seconds for the SVM-OVO (a), the $\mathcal{T}$-EFS-MCM-SGD (b), the $\mathcal{T}$-EFS-MCM-M-SGD (c), the SFS-LS-SVM (d) $\mathcal{T}$-EFS-LS-MCM-SGD (e), the $\mathcal{T}$-EFS-LS-MCM-M-SGD (f) for various quantization levels (integer and fraction bits) on the large datasets. 169
6.8 Heatmap of log2 mean training time + quantization time in seconds for the SVM-OVO (a), the $\mathcal{T}$-EFS-MCM-SGD (b), the $\mathcal{T}$-EFS-MCM-M-SGD (c), the SFS-LS-SVM (d) $\mathcal{T}$-EFS-LS-MCM-SGD (e), the $\mathcal{T}$-EFS-LS-MCM-M-SGD (f) for various quantization levels (integer and fraction bits) on the large datasets (cont.). 170
6.8 Heatmap of log2 mean training time + quantization time in seconds for the SVM-OVO (a), the $\mathcal{T}$-EFS-MCM-SGD (b), the $\mathcal{T}$-EFS-MCM-M-SGD (c), the SFS-LS-SVM (d) $\mathcal{T}$-EFS-LS-MCM-SGD (e), the $\mathcal{T}$-EFS-LS-MCM-M-SGD (f) for various quantization levels (integer and fraction bits) on the large datasets (cont.). 171

List of Tables

2.1 Datasets used throughout the thesis 26
2.2 Hyperparameter tuning range across algorithms 27
2.3 Notations used in the graphs 27
2.4 p-values for accuracies comparison across algorithms 31
2.5 p-values for run time comparison across algorithms 31

3.1 Datasets used throughout the thesis 53
3.2 Hyperparameter tuning range across algorithms 54
3.3 Notations used in the graphs 54
3.4 p-values for accuracies comparison across non least squares algorithms 56
3.5 p-values for nSV comparison across non least squares algorithms . . . 58
3.6 p-values for L_2 norm comparison across non least squares algorithms 58
3.7 p-values for run-time comparison across non least squares algorithms 58
3.8 p-values for accuracies comparison across least squares algorithms . . 62
3.9 p-values for nSV comparison across least squares algorithms 62
3.10 p-values for L_2 norm comparison across least squares algorithms . . . 62
3.11 p-values for run time (s) comparison across least squares algorithms . 63

4.1 Datasets used throughout the thesis 80
4.2 Hyperparameter tuning range across algorithms 80
4.3 Notations used in the graphs 82
4.4 p-values for accuracies comparison across non least squares algorithms 82
4.5 p-values for nSV comparison across non least squares algorithms . . . 83
4.6 p-values for L_2 norm comparison across non least squares algorithms 83
4.7 p-values for run time comparison across non least squares algorithms 83
4.8 p-values for accuracies comparison across least squares algorithms . . 88
4.9 p-values for nSV comparison across least squares algorithms 88
4.10 p-values for L_2 norm comparison across least squares algorithms . . . 88
4.11 p-values for run time comparison across least squares algorithms . . . 89
4.12 Image and text classification datasets and the number of features after processing 91
4.13 Test set accuracies on large non deep learning datasets 92
4.14 Training time on large non deep learning datasets 92
4.15 Test set accuracies on deep learning datasets 93
4.16 Training time on deep learning datasets 93

5.1 Datasets used throughout the thesis 107
5.2 Hyperparameter tuning range across algorithms 107
5.3 Notations used in the graphs 110
5.4 p-values for accuracies comparison across non least squares algorithms 116
5.5 p-values for nSV comparison across non least squares algorithms . . . 116

5.6 p-values for L_2 norm comparison across non least squares algorithms 117
5.7 p-values for training time comparison across non least squares algorithms . . . 117
5.8 p-values for accuracy comparison across least squares algorithms . . . 125
5.9 p-values for nSV comparison across least squares algorithms . . . 125
5.10 p-values for L_2 norm comparison across least squares algorithms . . . 126
5.11 p-values for training time comparison across least squares algorithms 126

6.1 Datasets used throughout the thesis . . . 138
6.2 Hyperparameter tuning range across algorithms . . . 138
6.3 Notations used in the graphs . . . 139
6.4 Small datasets summary . . . 140
6.5 large datasets summary . . . 140

List of Equations

1.1 Generalization Error . . . 3
1.2 Empirical Error . . . 3
1.3 PAC Learnability . . . 4
1.5 Growth function . . . 4
1.6 VC dimension . . . 4
1.6 VC Dimension Generalization Bound . . . 4
1.7 Simple VC bound . . . 5
1.8 SRM . . . 5
1.9 SVM hypothesis function . . . 5
1.11 Hinge loss . . . 5
1.13 VC dimension of canonical hyperplanes . . . 6
1.16 SVM primal . . . 6
1.18 SVM dual . . . 6
1.19 Kernel function . . . 7
1.20 Mercer's condition . . . 7
1.23 Reproducing Kernel Hilbert Space . . . 7
1.24 Representer Theorem . . . 7
1.25 VC dimension of kernel-based hypothesis . . . 8
1.27 ρ Margin loss . . . 8
1.29 Empirical Rademacher Complexity . . . 8
1.30 Rademacher Complexity . . . 9
1.33 SVM Rademacher Complexity based generalization bound . . . 9
1.37 MCM Primal . . . 10
1.41 MCM dual . . . 10
2.1 VC bound for gap tolerant classifiers . . . 15
2.2 MCM VC tight bound . . . 16
2.9 Linear MCM . . . 16
2.10 Linear hypothesis function . . . 16
2.14 Linear MCM modified . . . 17
2.25 kernel MCM modified . . . 18
2.29 kernel L_1 MCM modified . . . 18
2.36 Linear L_1 MCM modified . . . 19
2.40 RFF feature map . . . 20
2.41 Nystrom feature map . . . 21
2.43 FastFood feature map . . . 21
2.45 Non linear L_1 MCM modified . . . 22
2.71 Linear L_1 MCM modified convergence . . . 25
3.2 A vector in Empirical Feature Space (EFS) . . . 35
3.5 Hypothesis function in the EFS . . . 35
3.8 Canonical EFS . . . 35

3.9 VC dimension in the EFS . . . 36
3.10 EFS constraints . . . 36
3.11 kernel EFS constraints . . . 36
3.14 Kernel L_0 EFS-MCM . . . 37
3.17 Kernel L_1 EFS-MCM . . . 37
3.22 Tight VC bound on Kernel L_1 EFS-MCM . . . 37
3.33 Hoeffding's inequality . . . 38
3.55 Kernel L_0 EFS-MCM margin version . . . 40
3.61 Kernel L_1 EFS-MCM-M (margin version) . . . 41
3.67 LS-SVM . . . 42
3.68 Kernel LS-SVM optimization algorithm . . . 42
3.70 Kernel L_0 EFS-LS-MCM . . . 43
3.72 Kernel L_1 EFS-LS-MCM . . . 43
3.78 Kernel L_1 EFS-LS-MCM-M (margin version) . . . 43
3.83 Differential Renyi Entropy of order $q > 0$ . . . 44
3.83 Density estimation function . . . 44
3.84 Quadratic Renyi Entropy . . . 45
3.90 Large scale kernel L_1 EFS-MCM . . . 45
3.96 Large scale kernel L_1 EFS-MCM-M (margin version) . . . 46
3.100 Large scale kernel L_1 EFS-LS-MCM . . . 46
3.104 Large scale kernel L_1 EFS-LS-MCM-M (margin version) . . . 47
3.113 SFS-LS-SVM system of linear equations . . . 47
4.1 Multiclass hypothesis function . . . 66
4.2 Multiclass margin definition . . . 66
4.3 Empirical multiclass loss . . . 66
4.6 $\mathcal{T}$-EFS-MCM optimization problem . . . 66
4.15 Multiclass Rademacher generalization bound . . . 67
4.27 Bounding the Rademacher Complexity of $\Pi_1(\mathcal{F})$ . . . 69
4.48 $\mathcal{T}$-EFS-LS-MCM optimization problem . . . 71
4.51 $\mathcal{T}$-EFS-MCM-M optimization problem . . . 71
4.52 $\mathcal{T}$-EFS-LS-MCM-M optimization problem . . . 71
4.53 Large scale $\mathcal{T}$-EFS-MCM . . . 72
4.54 Large scale $\mathcal{T}$-EFS-LS-MCM . . . 72
4.55 Large scale $\mathcal{T}$-EFS-LS-MCM . . . 72
4.56 Large scale $\mathcal{T}$-EFS-LS-MCM-M . . . 72
5.2 Optimization problem types . . . 97
5.3 EFS optimization problem in the RKHS . . . 98
5.4 EFS optimization problem in the RKBS . . . 98
5.16 EFS optimization problem in the RKKS . . . 100
5.19 Krein space . . . 100
5.21 Associated Hilbert space . . . 100
5.23 $\mathcal{T}$-EFS-MCM loss function . . . 101
5.25 $\mathcal{T}$-EFS-LS-MCM loss function . . . 101
5.26 $\mathcal{T}$-EFS-MCM objective function . . . 101
5.27 $\mathcal{T}$-EFS-LS-MCM objective function . . . 101
5.28 $\mathcal{T}$-EFS-MCM-M objective function . . . 102
5.29 $\mathcal{T}$-EFS-LS-MCM-M objective function . . . 102
5.30 Krein space decomposition for margin term . . . 102

5.34 Eigen-decomposition of an indefinite kernel matrix 102
5.38 Large scale $\mathcal{T}$-EFS-MCM objective function 103
5.39 Large scale $\mathcal{T}$-EFS-LS-MCM objective function 103
5.40 Large scale $\mathcal{T}$-EFS-MCM-M objective function 104
5.41 Large scale $\mathcal{T}$-EFS-LS-MCM-M objective function 104
6.0 Signed bit representation of a scalar λ_i 130
6.0 3 bit representation of a scalar λ_i 130
6.2 Hypothesis function f^Q quantized version 131
6.3 Large scale $\mathcal{T}$-EFS-MCM objective function 132
6.4 Large scale $\mathcal{T}$-EFS-LS-MCM objective function 132
6.5 Large scale $\mathcal{T}$-EFS-MCM-M objective function 132
6.6 Large scale $\mathcal{T}$-EFS-LS-MCM-M objective function 132
6.6 Hinge loss function 133
6.6 L_1 Regularizer 133
6.6 Binary optimization problem for the EFS-LS-MCM-ILP 133
6.6 L_1 loss function 133
6.6 L_1 regularizer 133
6.6 Binary optimization problem for the EFS-LS-MCM-ILP 134
6.8 SVM Primal 134
6.10 SVM dual 134
6.12 Fixed size LS-SVM primal 135
6.13 FS-LS-SVM system of linear equations 135
6.14 SFS-LS-SVM OVO based objective function 135
6.14 SFS-LS-SVM OVO based objective function matrix notation 135
6.31 Difference in margin between the quantized and non-quantized versions . 137