Kernels: Regularization and Optimization

Cheng Soon Ong

A thesis submitted for the degree of
Doctor of Philosophy at
The Australian National University

April 2005
Except where otherwise indicated, this thesis is my own original work.

Cheng Soon Ong
12 April 2005
To my parents,
Ong Chee Lai and Ooi Eng Wah.
Acknowledgements

There are many people who have directly or indirectly made it possible for me to continue on this journey of discovery. At the risk of forgetting to mention some, I would like to sincerely thank the following people (in no particular order): The Australian Taxpayer, Michelle Moravec, Pamela Philips, Eric McCreath, Zoubin Ghahramani, Ulrike von Luxburg, Matthias Hein, Agnes Boskovitz, Edward Harrington, Evan Greensmith, Doug Aberdeen, S.V.N. Vishwanathan, Lydia Knüfing, Markus Hegland, Lai Weng Kin, John Shawe-Taylor and Bernhard Schölkopf.

I would especially like to thank my supervisors John W. Lloyd and Alexander J. Smola for all the support and advice that they have given me these past few years. They have made sure that I did not get too distracted by the subtle hedonistic pleasures of Canberra, and paved the way for me to have very productive research visits to the Gatsby Institute in University College London, Microsoft Research Cambridge, the Max Planck Institute in Tübingen and the University of Southampton. Stéphane Canu, whom I shared an office with for several highly productive months when he visited ANU, had a large influence on the later chapters of this thesis. I would also like to thank the co-authors of my conference publications and submissions Xavier Mary, Petra Philips and Bob Williamson for the pleasure for working with them and for their permission to use our joint work in this thesis. Large parts of Chapter 2 was published in Ong et al. [2003], and a shortened version of Chapter 3 was published in Ong and Smola [2003]. The main ideas of Chapter 4 was published in Ong et al. [2004]. Although I was the principal author on the above publications all the work was done jointly with my coauthors and it is impossible to disentangle each individual contribution.

I am indebted to the Australian National University Malaysian Alumni scholarship for supporting me financially during the period of my PhD, and Mimos Berhad who kindly allowed me to be on study leave all this time, hence allowing me to do my research without worrying about where my next meal will come from.

Last but not least, I thank my friends and family who have kept me sane and reminded me that there is more to life than a PhD.
Abstract

This thesis extends the paradigm of machine learning with kernels. This paradigm is based on the idea of generalizing an inner product between vectors to a similarity measure between objects. The kernel implicitly defines a feature mapping between the space of objects and the space of functions, called the reproducing kernel Hilbert space. There have been many successful applications of positive semidefinite kernels in diverse fields. Among the reasons for its success are a theoretically motivated regularization method and efficient algorithms for optimizing the resulting problems.

Since the kernel has to effectively capture the domain knowledge in an application, we study the problem of learning the kernel itself from training data. The proposed solution is a kernel on the space of kernels itself, which we called a hyperkernel. This provides a method for regularization via the norm of the kernel. We show that for several machine learning tasks, such as binary classification, regression and novelty detection, the resulting optimization problem is a semidefinite program. We solve the corresponding optimization problems using the same parameter settings across all problems, and demonstrate that we have further automated machine learning methods.

We observe that the restriction for kernels to be positive semidefinite can be removed. The non-positive kernels, called indefinite kernels, have corresponding functional theory, and define reproducing kernel Krein spaces. We derive machine learning problems with indefinite kernels and prove the representer theorem as well as generalization error bounds.

We provide theoretical and experimental evidence to support the idea of regularization by early stopping of conjugate gradient type algorithms. Conjugate gradient type algorithms are iterative methods that generate solutions in Krylov subspaces, and exhibit semi-convergence. We analyse the sequence of Krylov subspaces that determine the associated filter function on the spectrum of the inverse problem, and quantitatively investigate semi-convergence. These algorithms are then used for machine learning with indefinite kernels.
Contents

Acknowledgements vii
Abstract ix

1 Introduction 1
  1.1 Introduction to Machine Learning 1
  1.1.1 Kernels 3
  1.1.2 Regularization 5
  1.1.3 Optimization 6
  1.2 Contributions of this Thesis 7
  1.2.1 Learning the Kernel 7
  1.2.2 Learning with Indefinite Kernels 8
  1.2.3 Regularization by Early Stopping 8
  1.3 Summary of Contributions 9

2 Learning the Kernel 11
  2.1 Introduction 11
  2.2 Review of Methods for Model Selection 12
  2.2.1 Model Selection for Machine Learning 12
  2.2.2 Learning the Kernel 14
  2.3 Kernel Quality Functionals 14
  2.3.1 Empirical and Expected Quality 15
  2.3.2 Examples of $Q_{\text{emp}}$ 17
  2.4 Hyper Reproducing Kernel Hilbert Spaces 20
  2.4.1 Regularized Quality Functional 20
  2.4.2 A Bayesian Perspective 23
  2.5 Hyperkernels 26
  2.5.1 Power Series Construction 26
  2.5.2 Hyperkernels Invariant to Translation 28
  2.5.3 Explicit Expansion 29
  2.5.4 Summary of Hyper-RKHS 29
3 Machine Learning with Hyperkernels

3.1 Optimization Problem ............................................. 31
  3.1.1 Semidefinite Programming Formulations ..................... 32
  3.1.2 Examples of Hyperkernel Optimization Problems ........... 34
3.2 Experiments ....................................................... 37
  3.2.1 Low Rank Approximation ....................................... 39
  3.2.2 Classification Experiments .................................... 39
  3.2.3 Effect of $\lambda_Q$ and $\lambda_h$ on Classification Error .... 41
  3.2.4 Regression Experiments ..................................... 42
  3.2.5 Novelty Detection ............................................. 44
3.3 Summary .......................................................... 44

4 Learning with Indefinite Kernels .................................. 49

4.1 Why Non-Positive Kernels? ......................................... 49
4.2 Reproducing Kernel Krein Spaces ................................ 50
  4.2.1 Krein spaces ..................................................... 51
  4.2.2 From Krein spaces to Kernels ................................ 53
  4.2.3 From Kernels to Krein spaces ................................ 53
  4.2.4 Examples and Spectral Properties ........................... 54
4.3 Machine Learning in RKKS ....................................... 55
  4.3.1 Projection and Stabilization ................................ 56
  4.3.2 Application to general spline smoothing .................... 58
4.4 Generalization Bounds via Rademacher Average .................. 59
4.5 Spectrum of the Evaluation Operator ........................... 61
4.6 Summary of Learning in Krein spaces ............................. 65

5 Regularization by Early Stopping .................................. 67

5.1 What is Regularization? ........................................... 67
5.2 Krylov Subspace Algorithms ..................................... 69
  5.2.1 Iteration and Residual Polynomials ......................... 70
  5.2.2 Details of Conjugate Gradient Methods .................... 71
  5.2.3 Relation to Partial Least Squares ......................... 74
5.3 Filter Functions ................................................... 74
  5.3.1 Truncated Spectral Factorization ............................. 75
  5.3.2 Tikhonov Regularization ..................................... 75
  5.3.3 Steepest Descent .............................................. 76
  5.3.4 Krylov Subspace Methods ................................... 77
5.4 Semi-convergence of Minimal Residual ......................... 78
5.5 Conclusion ................................................................. 88

6 Applications of Indefinite Kernels 89
  6.1 Principal Component Analysis ................................. 90
  6.2 Partial Least Squares Regression ............................ 92
  6.3 Fisher Discriminant .............................................. 94
  6.4 Conclusion .......................................................... 97

7 Conclusion ............................................................... 99

A Derivations and Proofs for Hyperkernels 101
  A.1 Proof of Proposition 14 ........................................... 101
  A.2 Derivation of Hyperkernel Optimization ...................... 102
    A.2.1 \( L_1 \) SVM (C-parameterization) ....................... 103
    A.2.2 \( L_1 \) SVM (\( \nu \)-parameterization) ................. 105
    A.2.3 \( L_2 \) SVM (Lagrangian SVM) ......................... 108
    A.2.4 Singleclass SVM ........................................... 110
    A.2.5 \( \nu \)-Regression ........................................... 112

B Proof of Generalization Bounds .................................. 115

C Details of proof of semi-convergence 117
  C.1 Eigenvalue decomposition ....................................... 117
  C.2 Lemmas for Bounding the Residual ........................... 117
  C.3 Lemmas for Bounding the Error .............................. 121
List of Symbols

These are the symbols used in the thesis, and the page where they are first defined

$K$  
Gram Matrix of kernel function, $K_{ij} := k(x_i, x_j)$ where $x_i, x_j \in \mathcal{X}$, page 14

$R$  
Expected Risk, page 5

$R_{\text{emp}}(\mathcal{F})$  
Rademacher average of a set of functions $\mathcal{F}$, page 60

$T$  
evaluation functional, page 52

$\mathcal{K}$  
Hilbert space associated with Kreın space, page 51

$\mathbb{R}$  
the real numbers, page 3

$Z$  
Set of training data and labels, page 25

$\mathcal{X}$  
Space of input data, page 2

$X_{\text{train}}$  
Training data, page 2

$\det K$  
determinant of the matrix $K$, page 14

$Q_{\text{emp}}$  
Empirical Quality Functional, page 15

$R_{\text{emp}}$  
Empirical Risk Functional, page 15

$k(x, y, s, t)$  
Hyperkernel, page 21

$\mathcal{S}_k(z; G)$  
Krylov subspace of rank $k$, page 70

$\mathcal{Y}$  
Space of input labels, page 2

$Y_{\text{train}}, y$  
Training labels, page 2

$\lambda_i$  
ith eigenvalue, page 61

$\ell$  
Loss function, page 5

$Q$  
Expected Quality Functional, page 15

$R$  
Expected Risk Functional, page 15
<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Q_{reg}$</td>
<td>Regularised Quality Functional</td>
<td>21</td>
</tr>
<tr>
<td>$\mathcal{K}$</td>
<td>reproducing kernel Kreın space</td>
<td>52</td>
</tr>
<tr>
<td>$\sigma_i$</td>
<td>$i$th singular value</td>
<td>64</td>
</tr>
<tr>
<td>spec$A$</td>
<td>spectrum of the operator $A$</td>
<td>62</td>
</tr>
<tr>
<td>tr$K$</td>
<td>trace of the matrix $K$</td>
<td>14</td>
</tr>
<tr>
<td>$k(x_i, x_j)$</td>
<td>kernel</td>
<td>3</td>
</tr>
<tr>
<td>$p_k(A)$</td>
<td>residual polynomials of order $k$</td>
<td>71</td>
</tr>
<tr>
<td>$q_k(A)$</td>
<td>iteration polynomials of order $k$</td>
<td>71</td>
</tr>
</tbody>
</table>