CONVERGENCE ANALYSIS OF REGULARIZATION ALGORITHMS IN LEARNING THEORY

ABHISHAKE

DEPARTMENT OF MATHEMATICS
INDIAN INSTITUTE OF TECHNOLOGY DELHI
OCTOBER 2017

CONVERGENCE ANALYSIS OF REGULARIZATION ALGORITHMS IN LEARNING THEORY

by

ABHISHAKE

Department of Mathematics

Submitted

in fulfillment of the requirements of the degree of Doctor of Philosophy

to the

Indian Institute of Technology Delhi
October 2017

Dedicated to

My Family and Supervisor

Certificate

This is to certify that the thesis entitled **"Convergence Analysis of Regularization Algorithms in Learning Theory"** submitted by **Mr. Abhishake** to the Indian Institute of Technology Delhi, for the award of the Degree of **Doctor of Philosophy**, is a record of the original bona fide research work carried out by his under my supervision and guidance. The thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree or diploma.

Date:
Place: New Delhi

Dr. Sivananthan Sampath
Assistant Professor
Department of Mathematics
Indian Institute of Technology Delhi

Acknowledgements

There are many ways of expressing one's gratitude towards someone. Language is one such way by means of which one can express one's gratefulness and indebtedness to one's guide and benefactors. Reaching foremost milestones in one's life does not occur without the assistance of many great people and for each one of them, I am truly blessed and obliged.

I would like to express my heartfelt gratitude to my supervisor, Dr. Sivananthan Sampath, for his proficient supervision and research acumen. Above all and the most needed, he provided me unflinching encouragement and support in various ways. I am privileged to have the opportunity to work under his supervision. With his valuable suggestions, understanding, patience, enthusiasm, vivid discussions and above all with his friendly nature, he helped me in successfully completing my Ph.D. work. I am always being honored to have a teacher like him.

I am thankful to IIT Delhi authorities for providing me the necessary facilities and teaching assistantship for smooth completion of my research work. Special thanks to my SRC (Student Research Committee) Members: Prof. Niladri Chatterjee (Department of Mathematics), Prof. Harish Kumar (Department of Mathematics) and Prof. Shubhendu Bhasin (Department of Electrical Engineering) for generously sharing their time and knowledge. I express my sincere thanks to Prof. S. Dharmaraja (HOD) and DRC members for their kind support. I'm also grateful to all faculty members of Department of Mathematics, IIT Delhi, for their cooperation and support.

Words cannot completely express my love and gratitude to my family who have supported and encouraged me during this journey. I would like to thank my parents for their life-long support and sacrifices, which sustained my interest in research and motivated me towards the successful completion of this study. My sincerest thanks to my sisters Naincy, Roopsi, Deepika and brother Suraj who have supported me through my difficult times with their inspiration and continuous encouragement.

I will always remember the true friends who have walked with me on my path. There are too many to mention by name, yet too few to ever lose room in my heart.

Many thanks to my dear friends, especially of Debasis, Sanjay, Upendra, Abhilash, Nitya and Abhinav for times of laughter when I needed a reprieve from the intensity of research.

Finally, I thank the almighty God for the passion, strength, perseverance, and the resources to complete this study.

New Delhi Abhishake
August 2017

Abstract

The main goal of the thesis is to analyze the convergence issues of the regularization algorithms in learning theory. We study a theoretical analysis of the performance of one-parameter regularization and the multi-penalty regularization over the reproducing kernel Hilbert space. We discuss the error estimates of the regularization schemes under some prior assumptions for the joint probability measure on the sample space. We analyze the convergence rates of learning algorithms measured in the norm in reproducing kernel Hilbert space and in the norm in Hilbert space of square-integrable functions. The convergence issues for the learning algorithms are discussed in probabilistic sense by exponential tail inequalities.

A widely used approach for learning a function from empirical data is to minimize a regularization functional consists of an error term measuring the fitness of data and a smoothness term measuring the complexity of the function. The regularization scheme is usually referred as Tikhonov regularization or least-square regularization. We discuss the upper convergence rates of Tikhonov regularization and general regularization schemes under general source condition. Further, we address the minimum possible error for any learning algorithm under some prior assumptions. The concept of effective dimension is exploited to achieve the optimal convergence rates.

Manifold regularization is an approach which exploits the geometry of the marginal distribution. It can be naturally extended to the multi-task learning which exploits output-dependencies as well as enforce smoothness with respect to input data geometry. Further, the concept of vector-valued manifold regularization is also generalized to the multi-view manifold regularization. We propose a more general multi-penalty framework and establish the optimal convergence rates under the general smoothness assumption in vector-valued function setting.

The expansion of automatic data generation and acquisition bring data of huge size and complexity which raises challenges to computational capacities. In order to tackle these difficulties, various techniques are discussed for single-penalty regularization in the literature. Nyström subsampling is discussed for dealing with

big data in large scale kernel methods. We use this idea to efficiently implement the multi-penalty regularization algorithm. We obtain optimal convergence rates of multi-penalty regularization under general source condition for Nyström type subsampling.

In order to optimize the regularization functional, one of the crucial issue is to select regularization parameters to ensure good performance of the solution. The most widely used approach to select regularization parameters is discrepancy principle. We propose a new parameter choice rule "the balanced-discrepancy principle" and analyze the convergence of the scheme with the help of estimated error bounds.

We construct various regularized solutions corresponding to different choices of regularization schemes, hypothesis spaces (RKHSs), subsampling sizes and parameter choice rules. We study an aggregation approach based on the linear functional strategy and discuss a theoretical justification of the scheme.

We demonstrate the superiority of multi-parameter regularization over single-penalty regularization on a series of test samples. We provide an empirical analysis of the multi-view manifold regularization scheme based on the linear functional strategy for the challenging multi-class image classification and species recognition with attributes. Finally, we consider the NSL-KDD benchmark dataset from UCI machine learning repository for an empirical analysis of Nyström type subsampling.

सार

शोध प्रबंध का मुख्य लक्ष्य सीखने के सिद्धांत में नियमितकरण एल्गोरिदम के अभिसरण मुद्दों का विश्लेषण करना है। हम सिंगल-पैरामीटर और मल्टी-पैरामीटर नियमितकरण के व्यवहार का रेप्रोड्यूसिंग कर्नेल हिल्बर्ट स्पेस पर सैद्धांतिक विश्लेषण करते हैं। हम सैंपल स्पेस पर संयुक्त संभाव्यता के लिए कुछ पूर्व अवधारणाओं के तहत नियमितकरण योजनाओं के त्रुटि अनुमानों पर चर्चा करते हैं। हम लर्निंग एल्गोरिदम की अभिसरण दर को रेप्रोड्यूसिंग कर्नेल हिल्बर्ट स्पेस और स्क्वायर-इंटेग्रेबल फंक्शन्स के हिल्बर्ट स्पेस की नॉर्म में विश्लेषित करते हैं । लर्निंग एल्गोरिदम के लिए अभिसरण मुद्दों पर घातीय पूंछ असमानताओं द्वारा संभाव्य अर्थों में चर्चा की जाती है।

आनुभविक डेटा से एक फ़ंक्शन को सीखने के लिए एक व्यापक रूप से इस्तेमाल किया जाने वाला दृष्टिकोण नियमित फंक्शनल को न्यूनतम करना है जिसमें डेटा की फिटनेस को मापने और फ़ंक्शन की जटिलता को मापने के लिए त्रुटि पद होते है। नियमितकरण योजना को आमतौर पर टिखोनोव नियमितकरण या लीस्ट-स्क्वायर नियमितकरण के रूप में जाना जाता है। हम व्यापक स्रोत शर्त के तहत टिखोनोव नियमितकरण और व्यापक नियमितकरण योजनाओं की ऊपरी अभिसरण दर पर चर्चा करते हैं। इसके अलावा, हम कुछ पूर्व अवधारणाओं के तहत किसी भी लर्निंग एल्गोरिदम के लिए न्यूनतम संभव त्रुटि को संबोधित करते हैं। सर्वोत्तम अभिसरण दर प्राप्त करने के लिए प्रभावी आयाम की अवधारणा का उपयोग किया जाता है।

मैनिफ़ोल्ड नियमितकरण एक दृष्टिकोण है जो मार्जिनल संभाव्यता के वितरण की ज्यामिति का उपयोग करता है। इसे स्वाभाविक रूप से मल्टी-टास्क लर्निंग के लिए विस्तारित किया जा सकता है जो आउटपुट-निर्भरता का लाभ उठाता है और इनपुट डेटा ज्यामिति में मसृणता को लागू करता है। इसके अलावा, वेक्टर-वैल्यूड मैनिफ़ोल्ड नियमितकरण की अवधारणा को बहु-दृश्य मैनिफ़ोल्ड नियमितकरण के लिए विस्तारित किया गया है। हम एक व्यापक मल्टी-पेनल्टी ढांचे का प्रस्ताव करते हैं और स्केलर-वैल्यूड फ़ंक्शन सेटिंग में व्यापक अवधारणाओं के तहत सर्वोत्तम अभिसरण दर स्थापित करते हैं।

स्वचालित डेटा उत्पादन और अधिग्रहण का विस्तार, विशाल आकार और जटिलता का डेटा लाता है जो कम्प्यूटेशनल क्षमताओं के लिए चुनौतियां बढ़ाता है। इन कठिनाइयों से निपटने के लिए, साहित्य में सिंगल-पेनल्टी नियमितकरण के लिए विभिन्न तकनीकों की चर्चा की जाती है। बड़े पैमाने की कर्नेल पद्धतियों में बड़े डेटा से निपटने के लिए निस्ट्रॉम सबसेप्लिंग पर चर्चा की जाती है। हम इस विचार का उपयोग मल्टी-पेनल्टी नियमितकरण एल्गोरिथ्म को प्रभावी ढंग से लागू करने के लिए करते हैं। हम व्यापक स्रोत शर्त के तहत निस्ट्रॉम सबसेप्लिंग के लिए मल्टी-पेनल्टी नियमितकरण की सर्वोत्तम अभिसरण दर प्राप्त करते हैं।

नियमितकरण फंक्शनल को ऑप्टिमाइज़ करने में अच्छे प्रदर्शन को सुनिश्चित करने के लिए नियमितकरण पैरामीटर चुनना एक महत्वपूर्ण समस्या है। नियमितकरण पैरामीटर का चयन करने के लिए सबसे अधिक व्यापक रूप से उपयोग किया जाने वाला दृष्टिकोण विसंगति सिद्धांत है। हम एक नया पैरामीटर चयन नियम "संतुलित-विसंगति सिद्धांत" का प्रस्ताव करते हैं और अनुमानित त्रुटि सीमा की सहायता से योजना के अभिसरण का विश्लेषण करते हैं।

हम नियमितकरण योजनाओं, हाइपोथिसिस स्पेसेस (आरकेएचएस), सबसेप्लिंग माप और पैरामीटर नियमों के विभिन्न विकल्पों के अनुरूप विभिन्न नियमित आगणक तैयार करते हैं। हम लीनियर फंक्शनल स्ट्रेटेजी के आधार पर एकत्रीकरण दृष्टिकोण का अध्ययन करते हैं और योजना के सैद्धांतिक औचित्य पर चर्चा करते हैं।

हम टेस्ट नमूने की एक श्रृंखला के आधार पर सिंगल-पेनल्टी नियमितकरण पर मल्टी-पेनल्टी नियमितकरण की श्रेष्ठता का प्रदर्शन करते हैं। हम चुनौतीपूर्ण बहु-श्रेणी की छवि वर्गीकरण और विशेषताओं के साथ प्रजातियों की पहचान के लिए लीनियर फंक्शनल स्ट्रेटेजी के आधार पर बहु-दृश्य मैनिफ़ोल्ड नियमितकरण योजना का एक अनुभवजन्य विश्लेषण प्रदान करते हैं। अंत में, हम एनएसएल-केडीडी बेंचमार्क डेटासेट पर निस्ट्रॉम टाइप सबसेप्लिंग के लिए एक अनुभवजन्य विश्लेषण करते है।

Contents

Certificate **i**

Acknowledgements **iii**

Abstract **v**

List of Figures **ix**

List of Tables **xi**

List of Notations **xiii**

1 Introduction **1**

1.1 Learning from examples 2

1.2 Reproducing kernel Hilbert space (RKHS) 4

1.3 Prior assumptions 7

1.3.1 Spectral theoretical parameter: Effective dimension 9

1.4 Convergence analysis of one-parameter regularization 10

1.5 Manifold regularization 17

1.6 Nyström type subsampling in Tikhonov regularization 19

1.7 Parameter choice rules 22

1.8 Aggregation approach based on linear functional strategy 23

2 Optimal rates for one-parameter general regularization **25**

2.1 Upper rates for Tikhonov regularization 25

2.2 Upper rates for general regularization schemes 33

2.3 Lower rates for general learning algorithms 41

2.4 Individual lower rates 47

3 Optimal rates for multi-penalty regularization **53**

3.1 Convergence rates for multi-penalty regularization 53

3.2 Error estimation of linear functionals 65
3.3 Optimal rates for multi-penalty regularization 71

4 Optimal rates for multi-penalty regularization based on Nyström type subsampling 75
4.1 Convergence analysis 76

5 Parameter choice rules for multi-penalty regularization 87
5.1 Balanced-discrepancy principle for multi-penalty regularization 88
5.2 Numerical Realization 94

6 A linear functional strategy to aggregate regularized solutions 101
6.1 An aggregation approach for multi-penalty regularization 102
6.2 Numerical Realization 106
6.2.1 The academic example 106
6.2.2 Caltech-101 dataset 109
6.2.3 NSL-KDD dataset 117

Bibliography 123

Bio-Data 133

List of Figures

5.1 Sample error vs Approximation error 88

5.2 The target function f_ρ (red line) and its estimator $f_{\mathbf{z},\lambda_0}$ (blue line) based on the balancing principle with $\lambda_0 = 1.4835 \times 10^{-3}$ and the empirical data $\mathbf{z}_{16}$ (red stars). 95

5.3 (a) The estimator $f_{\mathbf{z},\lambda_0}$ (blue line) based on the balancing principle with $\lambda_0 = 5.9855 \times 10^{-4}$, (b) The estimator $f_{\mathbf{z},\lambda_1}$ (blue line) based on discrepancy principle with $\lambda_1 = 3.4994 \times 10^{-4}$. The target function f_ρ (red line) and the empirical data $\mathbf{z}_{16}$ (red stars). 95

5.4 (a) The target function f_ρ (red line) and its estimator $f^{(1)}_{\mathbf{z},\lambda}$ (blue line) based on the balanced-discrepancy principle with $\lambda_0 = 5.9855 \times 10^{-4}$, $\lambda_1 = 3.4518 \times 10^{-4}$ and the empirical data $\mathbf{z}_{16}$ (red stars), (b) The discrepancy curve is shown with the balanced-discrepancy parameter choice (red star). 96

5.5 Figures show relative errors in infinity norm (a), $||\cdot||_m$-empirical norm (b) and $||\cdot||_{\mathcal{H}}$-norm (c) corresponding to 100 test problems with samples according to (5.18) with $\delta = 0.02$ for all estimators. . . 97

5.6 The figures show the decision surfaces generated with two labeled samples (red star) by single-penalty regularizer (a) based on balancing principle ($\lambda_A = 6 \times 10^{-9}$) and manifold regularizer (b) based on balanced-discrepancy parameter choice $\lambda_A = 6 \times 10^{-9},\ \lambda_I = 0.0077$. . 99

6.1 The constructed solution $f^2_{\mathbf{z}}$ (blue line) based on LFS-$\mathscr{L}^2_{\rho_X}$, the ideal estimator f_ρ (red line), the approximants $\{f_{\mathbf{z},\lambda^i}\}_{i=1}^4$ (green lines) and the empirical data $\mathbf{z}_{16}$ (red stars). 107

6.2 Figures show relative errors in $||\cdot||_{\mathcal{H}}$-norm (a), $||\cdot||_m$-empirical norm (b) and infinity norm (c) corresponding to 100 test problems with samples according to (5.18) with $\delta = 0.02$ for all estimators. 108

List of Tables

1.1 Illustration of convergence rates for different regularization schemes . 21

5.1 Statistical performance interpretation of single-penalty regularizer $f_{\mathbf{z},\lambda_0}$. . 98
5.2 Statistical performance interpretation of single-penalty regularizer $f_{\mathbf{z},\lambda_1}$. . 98
5.3 Statistical performance interpretation of multi-penalty regularizer $f^{(1)}_{\mathbf{z},\lambda}$. . . 98
5.4 Statistical performance interpretation of multi-penalty regularizer $f^{(2)}_{\mathbf{z},\lambda}$. . . 98
5.5 Statistical performance interpretation of multi-penalty regularizer $f^{(3)}_{\mathbf{z},\lambda}$. . . 98
5.6 Statistical performance interpretation of single-penalty ($\lambda_I = 0$) and multi-penalty regularizers of the functional (3.1). 99

6.1 Statistical performance interpretation of the estimators in $\mathcal{H}$-norm. 108
6.2 Statistical performance interpretation of the estimators in empirical $\mathscr{L}^2_{\rho_X}$-norm. 108
6.3 Statistical performance interpretation of the estimators in sup-norm. . . . 108
6.4 Results of single-view learning on Caltech-101 using each feature and based on LFS-$\mathscr{L}^2_{\rho_X}$. 112
6.5 Results of multi-view learning on Caltech-101 using feature on each level of spatial pyramid and based on LFS-$\mathscr{L}^2_{\rho_X}$. 112
6.6 Results of multi-view learning on Caltech-101 using feature on each level of spatial pyramid and based on LFS-$\mathscr{L}^2_{\rho_X}$ with optimal choice of combination operator. 112
6.7 Results of multi-view learning on Caltech-101 using feature on each level of spatial pyramid and based on LFS-$\mathscr{L}^2_{\rho_X}$ with the balanced-discrepancy principle parameter choice. 113
6.8 Results of multi-view learning on Caltech-101 using feature on each level of spatial pyramid and based on LFS-$\mathscr{L}^2_{\rho_X}$ with the balanced-discrepancy principle parameter choice and optimal combination operator. 113

6.9 Performance of multi-view learning estimators based on Nyström subsampling and the aggregated solution based on LFS-$\mathscr{L}^2_{\rho_X}$ using 5 labeled and 5 unlabeled images per class from Caltech-101 data set. 116
6.10 Performance of multi-view learning estimators based on Nyström subsampling and the aggregated solution based on LFS-$\mathscr{L}^2_{\rho_X}$ using 10 labeled and 5 unlabeled images per class from Caltech-101 data set. 116
6.11 Performance of multi-view learning estimators based on Nyström subsampling and the aggregated solution based on LFS-$\mathscr{L}^2_{\rho_X}$ with optimal combination operator using 5 labeled and 5 unlabeled images per class from Caltech-101 data set. 116
6.12 Performance of multi-view learning estimators based on Nyström subsampling and the aggregated solution based on LFS-$\mathscr{L}^2_{\rho_X}$ with optimal combination operator using 10 labeled and 5 unlabeled images per class from Caltech-101 data set. 116
6.13 Statistical performance of various estimators on different sub-datasets (Folds) of NSL-KDD dataset using random subsampling 50 times. 119
6.14 Statistical performance of various estimators over all 9 sub-datasets (Folds) of NSL-KDD dataset. 120
6.15 Performance comparison of the proposed model with some existing research work on NSL-KDD dataset. 120

List of Notations

Symbol	Meaning
$\mathbb{N}$	The set of natural numbers
$\mathbb{R}$	The set of real numbers
$\mathbb{R}^n$	The n-dimensional Euclidean space
X	The input space
Y	The output space
Z	The sample space $X \times Y$
$\mathbf{x}$	m-tuple of inputs $(x_1, \ldots, x_m)$
$\mathbf{y}$	m-tuple of outputs $(y_1, \ldots, y_m)$
$\mathbf{z}$	m-tuple of samples $((x_1, y_1), \ldots, (x_m, y_m))$
ρ	The joint probability measure on the sample space Z
ρ_X	The marginal probability measure on X
$\rho(y\|x)$	The conditional probability measure on Y with respect to x
$\mathscr{L}^2_{\rho_X}$	Hilbert space of square integrable functions with respect to ρ_X
$\|\|\cdot\|\|_\rho$	The norm in $\mathscr{L}^2_{\rho_X}$, that is, $\|\|\cdot\|\|_{\mathscr{L}^2_{\rho_X}} = \{\int_X \|\|f(x)\|\|^2_Y d\rho_X(x)\}^{1/2}$
f_ρ	The regression function
$f_{\mathcal{H}}$	The target function with respect to the Hilbert space $\mathcal{H}$
$f_{\mathbf{z}}$	An estimator of the regression function based on samples $\mathbf{z}$
$f_{\mathbf{z},\lambda}$	The regularized estimator of the regression function based on samples $\mathbf{z}$ with regularization parameter λ
$f^l_{\mathbf{z}}$	The estimator of the regression function based on samples $\mathbf{z}$ corresponding to learning algorithm l

Symbol	Meaning
$S_{\mathbf{x}}$	The sampling operator
ϕ	The continuous increasing index function
$\mathcal{N}(\lambda)$	The effective dimension
E	The expectation
$Prob$	The probability
$\mathcal{L}(\mathcal{H})$	The set of bounded linear operators on $\mathcal{H}$
$\|\|\cdot\|\|_{\mathcal{L}(\mathcal{H})}$	The operator norm
$\|\|\cdot\|\|_{\mathcal{L}_2(\mathcal{H})}$	The Hilbert-Schmidt norm
$\mathbf{x}_s$	A subsample of the input points $\mathbf{x}$
$\mathcal{H}^{\mathbf{x}_s}$	The linear span of K_{x_i}'s corresponding to $\mathbf{x}_s$
$P_{\mathbf{x}_s}$	The orthogonal projection operator with range $\mathcal{H}^{\mathbf{x}_s}$
$Co\{x_i\}_{i=1}^m$	The convex hull of a finite point set $\{x_i\}_{i=1}^m$
$\square$	end of a proof