

Efficient and Robust Black-box Integral-approximation and Optimization

Yueming Lyu

Faculty of Engineering and Information Technology
University of Technology Sydney

A thesis submitted for the degree of
Doctor of Philosophy

September 2021

Certificate of Original Authorship

I hereby declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise reference or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis. The contents of this dissertation are original and have not been submitted in whole or in part for qualifications at any other academic institution. This research is supported by the Australian Government Research Training Program.

Signature:	Production Note: Signature removed prior to publication.
Date:	01-October-2021

I would like to dedicate this thesis to my loving parents, who support me all the time.

Acknowledgements

First, I would like to express my deepest gratitude to my supervisor, Prof. Ivor W. Tsang, who gave me the research opportunity in his group. My research journey was not easy. I was depressed and suffered a lot. It was Prof. Ivor W. Tsang who provided me an opportunity to continue research on machine learning. Prof. Ivor W. Tsang always supported me to be patient to do in-depth analysis instead of shallow work. His insightful guidance helped me to focus on the key reasons to solve the problems. Prof. Ivor W. Tsang encouraged me to explore the research area that I am interested. He guided me to know how to think in both theoretical and empirical views of research. Prof. Ivor W. Tsang taught me how to divide and conquer problems and hierarchically solve them. His critical thinking and broad view gave me insightful guidance and let me know how a good researcher should be. Prof. Ivor W. Tsang's advice and support improved my skills and confidence to be an independent researcher. It is a great fortune for me to be a Ph.D. student under his supervision.

Second, I would like to thank many friends who have supported me. I want to thank Dr. Jiangchao Yao and Dr. Yuangang Pan for discussing both research and life. It was a memorable time for waiting for the subway and discussing research together. I appreciate the time with them and Xiaowei Zhou, Xingrui Yu, Xu Chen, Yan Zhang, Jin Li, Yaxin shi, Jinliang Deng, Yinghua Yao for discussion and lunch together. I am so grateful for such a relaxed life and beautiful experience in Sydney. I would also like to thank Dr. Yuan Yuan, Dr. Yanbin Liu, Dr. Xin Yu, Dr. Xiaolin Zhang, Fan Ma, Tianqi Tang, Guangrui Li, and Guang Li for their support and discussion. I would thank Dr. Peng Sun and Dr. Li Shen for the discussion and for working together. I may miss mentioning many others, but I thank them for their help.

Finally, I would like to dedicate this thesis to my dearest parents and family. Thank you so much for being there, supporting me all the time.

Abstract

Black-box optimization and black-box integral approximation are important techniques for machine learning, industrial design, and simulation in science. This thesis investigates black-box integral approximation and black-box optimization by considering the closed relationship between them. For integral approximation, we develop a simple closed-form rank-1 lattice construction method based on group theory. Our method reduces the number of distinct pairwise distance values to generate a more regular lattice. Furthermore, we investigate structured points set for integral approximation on hyper-sphere. Our structured point sets can serve as a good initialization for black-box optimization. Moreover, we propose stochastic black-box optimization with implicit natural gradients for black-box optimization. Our method is very simple and has only the step-size hyper-parameter. Furthermore, we develop a batch Bayesian optimization algorithm from the perspective of frequentist kernel methods, which is powerful for low-dimensional black-box optimization problems. We further apply our structured integral approximation techniques for kernel approximation. In addition, we develop structured approximation for robust deep neural network architecture, which results in an elegant and simple architecture that preserves optimization properties. Moreover, we develop adaptive loss as a tighter upper bound approximation for expected 0-1 risk, robust and trainable with SGD.

Contents

1 Introduction	1
1.1 Black-box optimization	1
1.1.1 Sampling based Optimization	1
1.1.2 Bayesian Optimization	3
1.2 Black-box Integral Approximation	5
1.3 Connection Between Black-box Integral Approximation and Black-box Optimization	7
1.4 Integral Approximation for Kernel Methods and Deep Learning	8
1.4.1 Integral Approximation for Kernel Approximation	8
1.4.2 Structured Integral Approximation for Robust Neural Network Architecture	10
1.4.3 Adaptive Loss As A Tighter Upper Bound Approximation of Expected 0-1 Risk	12
1.5 Research Objectives	13
1.6 Publications	13
2 Related Works	15
2.1 Black-box Optimization	15
2.1.1 Bayesian Optimization	16
2.1.2 Sampling-based Derivative-free Optimization	17
2.2 Black-box Integral Approximation	18
2.3 Integral Approximation for Kernel Methods and Deep Learning	19
2.3.1 Kernel Approximation	19
2.3.2 Deep Learning Theory from Optimization and Kernel Perspective	20
2.3.3 Robust Deep Learning under Label Noise	21

3	Implicit Natural Gradient Optimization	25
3.1	Chapter Abstract	25
3.2	Optimization with Exponential-family Sampling	25
3.3	Implicit Natural Gradient	27
3.4	Update Rule for Gaussian Sampling	28
3.4.1	Stochastic Update	29
3.4.2	Mean field approximation for acceleration	31
3.4.3	Direct Update for μ and Σ	31
3.5	Optimization for Discrete Variable	33
3.6	Convergence Rate	34
3.7	Empirical Evaluation	36
3.7.1	Evaluation on synthetic continuous test benchmarks	36
3.7.2	Evaluation on RL test problems	38
3.7.3	Evaluation on discrete test problems	39
3.8	Summary	40
4	Batch Bayesian Optimization	42
4.1	Chapter Abstract	42
4.2	Problem Setup	42
4.3	BO in Noise-Free Setting	44
4.3.1	Sequential Selection in Noise Free Setting	44
4.3.2	Batch Selection in Noise-Free Setting	45
4.4	BO in Perturbation Setting	47
4.4.1	Sequential Selection in Perturbation Setting	47
4.4.2	Batch Selection in Perturbation Setting	48
4.5	Robust Initialization for BO	49
4.6	Fast Rank-1 Lattice Construction	52
4.6.1	The rank-1 lattice construction given a base vector	52
4.6.2	The separate distance of a rank-1 lattice	53
4.6.3	Searching the rank-1 lattice with maximized separate distance	54
4.6.4	Comparison of minimum distance generated by different methods	54
4.6.5	Comparison between lattice points and random points	55
4.7	Experiments	55
4.7.1	Comparison with Bull's Non-adaptive Batch Method	55
4.7.2	Empirical Evaluation on Synthetic Benchmark Problems	56

4.7.3	Empirical Evaluation on Hyperparameter tuning of Neural Network	57
4.7.4	Empirical Evaluation on Robot Pushing Task	59
4.8	Summary	60
5	Subgroup-based Rank-1 Lattice Quasi-Monte Carlo	61
5.1	Chapter Abstract	61
5.2	Background of Lattice	61
5.2.1	The Lattice	62
5.2.2	The separating distance of a lattice	62
5.3	Subgroup-based Rank-1 Lattice	63
5.3.1	Construction of the Generating Vector	63
5.3.2	Regular Property of Rank-1 Lattice	65
5.4	QMC for Kernel Approximation	66
5.5	Experiments	67
5.5.1	Evaluation of the minimum distance	67
5.5.2	Integral approximation	69
5.5.3	Kernel approximation	69
5.5.4	Approximation on Graphical Model	70
5.6	Subgroup-based QMC on Sphere \mathbb{S}^{d-1}	72
5.7	QMC for Generative models	74
5.8	Generative Inference for CycleGAN	75
5.9	Summary	76
6	Spherical Structured Feature Maps for Kernel Approximation	79
6.1	Chapter Abstract	79
6.2	Background of Kernel Approximation	79
6.2.1	Random Feature Maps	79
6.2.2	Discrete Riesz s-energy	81
6.3	Spherical Structured Feature Maps	82
6.3.1	Feature Maps for Shift and Rotation Invariant Kernels	82
6.3.2	Feature Maps for \mathbf{b}^{th} -order Arc-cosine Kernels	84
6.4	Design of Matrix \mathbf{U}	85
6.4.1	Structure of Matrix \mathbf{U}	86
6.4.2	Minimize the Discrete Riesz s-energy	86
6.5	Fast Feature Maps Construction	89
6.6	Empirical Studies	91

6.6.1	Convergence and Speedup	91
6.6.2	Approximation Accuracy	92
6.7	Summary	94
7	Neural Optimization Kernel: Towards Robust Deep Learning	95
7.1	Chapter Abstract	95
7.2	Neural Optimization Kernel	95
7.3	Structured Approximation	98
7.3.1	Convergence Rate for Finite Dimensional Approximation Prob-	
	lem	99
7.3.2	Learning Parameter \mathbf{R}	100
7.3.3	Kernel Approximation	102
7.4	Functional Optimization	103
7.5	Rademacher Complexity and Generalization Bound	104
7.6	Experiments	105
7.6.1	Empirical Evaluation on Classification under Gaussian Noise	
	Perturbation	106
7.6.2	Empirical Evaluation on Classification under Laplace Noise Per-	
	turbation	108
7.6.3	Empirical Evaluation on Classification with Adversarial Pertur-	
	bation	109
7.7	Summary	109
8	Curriculum Loss for Robust Deep Learning	111
8.1	Chapter Abstract	111
8.2	Curriculum Loss	111
8.2.1	Robustness of 0-1 loss against label corruption	112
8.2.2	Tighter upper bounds of the 0-1 Loss	113
8.2.3	Noise Pruned Curriculum Loss	117
8.3	Empirical Study	119
8.3.1	Evaluation of Robustness against Label Corruption	119
8.3.2	More experiments with different network architectures	122
8.3.3	Impact of Misspecified Estimation of Noise Rate ϵ	124
8.4	Summary	125
9	Conclusion	127

10 Appendix	128
10.1 Proof of Theorem 2	128
10.2 Proof of Theorem 3	128
10.3 Proof of Theorem 4	129
10.4 Proof of Theorem 5	137
10.5 Variance Reduction	139
10.6 Proof of Updating Theorem	142
10.7 Proof of Gradient and Hessian Theorem	143
10.8 Discrete Update	144
10.9 Proof of Theorem 6	145
10.10 Proof of Theorem 7	147
10.11 Proof of Theorem 8	151
10.12 Proof of Theorem 9	153
10.13 Proof of Theorem 10	156
10.14 Proof of Theorem 11	157
10.15 Proof of Theorem 12	157
10.16 Proof of Corollary 2	157
10.17 Proof of Theorem 13	158
10.18 Proof of Theorem 14	159
10.19 Proof of Corollary 1	161
10.20 Proof of Proposition 1	163
10.21 Proof of Theorem 22	163
10.22 Proof of Theorem 21	170
10.23 Proof of Theorem 23	172
10.24 Proof of Theorem 24	173
10.25 Proof of Theorem 25	176
10.26 A Better Diagonal Random Rotation for SSF	180
10.27 Rademacher Complexity	181
10.28 Generalization Bound	185
10.29 Rademacher Complexity and Generalization Bound for General Structured Neural Network Family	186
10.30 Explanation of Theorem 28 for robust learning	191
10.31 Proof of Theorem 29	193
10.32 Proof of Corollary 4	193
10.33 Proof of Partial Optimization Theorem (Theorem 31)	194
10.34 Proof of Proposition 2	195

10.35 Proof of Theorem 30	196
10.36 Proof of Corollary 5	197
10.37 Multi-Class Extension	198
10.38 Evaluation of Efficiency of the Proposed Soft-hinge Loss	198

List of Figures

3.1	Mean value of $f(\mathbf{x})$ in \log_{10} scale over 20 independent runs for 100-dimensional problems.	37
3.2	Average Reward over 5 independent runs on benchmark RL environments	39
3.3	Mean value of regret over 10 independent runs for different dimensional discrete optimization problems	40
4.1	Lattice Points and Random Points on $[0, 1]^2$	51
4.2	More Lattice Points and Random Points on $[0, 1]^2$	56
4.3	The mean value of simple regret over 30 runs on Rosenbrock and Ackley function	58
4.4	The mean value of simple regret on network tuning task and robot pushing task.	58
4.5	The mean value of simple regret for different algorithms over 30 runs on different test functions	59
5.1	Mean approximation error over 50 independent runs. error bars are with in $1 \times \text{std.}$	68
5.2	Relative Mean and Max Reconstruction Error for Gaussian, Zero-order and First-order Arc-cosine Kernel on DNA dataset. Error bars are within $1 \times \text{std.}$	70
5.3	Relative Mean and Max Reconstruction Error for Gaussian, Zero-order and First-order Arc-cosine Kernel on SIFT1M dataset.	71
5.4	Mean approximation error over 50 independent runs. Error bars are with in $1 \times \text{std}$	72
5.5	Illustration of the generated images from models trained with subgroup rank-1 lattice sampling, Monte-Carlo sampling, and Standard version of CycleGAN.	77

5.6	Illustration of the generated images from models trained with subgroup rank-1 lattice sampling, Monte-Carlo sampling, and Standard version of CycleGAN.	78
6.1	Convergence of the Logarithmic Energy	91
6.2	Speedup of the Feature Maps Construction	92
6.3	Relative Mean and Max Reconstruction Error for Gaussian, Zero-order and First-order Arc-cosine Kernel on MNIST	93
7.1	Mean test accuracy \pm std over 5 independent runs under Gaussian noise with DenseNet backbone	106
7.2	Mean test accuracy \pm std over 5 independent runs under Gaussian noise with ResNet backbone	107
7.3	Mean test accuracy \pm std over 5 independent runs under Laplace noise with DenseNet backbone	107
7.4	Mean test accuracy \pm std over 5 independent runs under Laplace noise with ResNet backbone	108
7.5	Mean test accuracy \pm std over 5 independent runs on CIFAR10/CIFAR100 dataset under FGSM adversarial attack for DenseNet and ResNet backbone.	109
8.1	Test accuracy and label precision vs. number of epochs on MNIST dataset.	120
8.2	Test accuracy and label precision vs. number of epochs on CIFAR10 dataset.	121
8.3	Test accuracy and label precision vs. number of epochs on CIFAR100 dataset.	122
8.4	Test accuracy vs. number of epochs on MNIST dataset.	123
8.5	Test accuracy vs. number of epochs on CIFAR100 dataset.	124
8.6	Test accuracy (%) on Tiny-ImageNet dataset with symmetric noise	126
8.7	Test accuracy vs. number of epochs on CIFAR10 dataset.	126
10.1	Training/Test accuracy for soft and hard hinge loss with different optimizer on CIFAR100	199

List of Tables

3.1 Test functions	37
4.1 Minimum distance ($2\rho_X$) of 1,000 lattice points in $[0, 1]^d$ for $d = 10$, $d = 20$, $d = 30$, $d = 40$ and $d = 50$.	54
4.2 Minimum distance ($2\rho_X$) of 2,000 lattice points in $[0, 1]^d$ for $d = 10$, $d = 20$, $d = 30$, $d = 40$ and $d = 50$.	54
4.3 Minimum distance ($2\rho_X$) of 3,000 lattice points in $[0, 1]^d$ for $d = 10$, $d = 20$, $d = 30$, $d = 40$ and $d = 50$.	55
4.4 Benchmark functions	57
5.1 Minimum l_1 -norm-based toroidal distance of rank-1 lattice constructed by different methods.	67
5.2 Minimum l_2 -norm-based toroidal distance of rank-1 lattice constructed by different methods.	67
5.3 Mutual coherence of points set constructed by different methods. Smaller is better.	74
7.1 Regularizers and Proximal Operators	96
8.1 Test accuracy(%) of DenseNet on CIFAR10 and CIFAR100.	123
8.2 Test accuracy(%) of DenseNet on CIFAR10 and CIFAR100 with se- mantic noise.	124
8.3 Test accuracy(%) of DenseNet on CIFAR10 with open-set noise.	125
8.4 Average test accuracy of NPCL with different ϵ on MNIST over last ten epochs	125

List of Algorithms

1	INGO	30
2	Fast INGO-u	31
3	INGOstep	33
4	General Framework	34
5	Sequential Noise-free Algorithm	45
6	Batch Noise-free Algorithm	46
7	Sequential Optimization with Perturbation	47
8	Batch Optimization with Perturbation	49
9	Greedy Batch Optimization	50
10	Rank-1 Lattice Construction	52
11	Rank-1 Lattice Construction with Successive Coordinate Search (SCS)	53
12	Coordinate Index Selection	89
13	Partial Optimization	116
14	Training with Batch Noise Pruned Curriculum Loss	119