

When Robust Machine Learning Meets Noisy Supervision: Mechanism, Optimization and Generalization

Bo Han

Centre for Artificial Intelligence
University of Technology Sydney

A thesis submitted for the degree of

Doctor of Philosophy

May, 2019

CERTIFICATE OF ORIGINAL AUTHORSHIP

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as part of the collaborative doctoral degree and/or fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This research is supported by an Australian Government Research Training Program.

Production Note:

Signature of Student: Bo Han

Signature removed prior to publication.

Date: 31/May/2019

I would like to dedicate this thesis to my loving parents and family.

Acknowledgements

I had a wonderful journey at University of Technology Sydney (UTS) for pursuing my PhD degree in the past four years. I am deeply grateful to all those who helped me during the writing of this thesis.

First of all, I would like to express my foremost and deepest gratitude to my advisor, Prof. Ivor W. Tsang, for his visionary, inspiring, and insightful guidance during my PhD period. Honestly speaking, I knew quite little about genuine research when I began my PhD study in UTS. He taught me how to find interesting ideas, how to develop solid algorithms, and how to write papers all from scratch. Besides his knowledge, his research taste encourages me towards the state-of-the-art algorithms, while his work ethic motivates me to become tough and strong. Inspiring by his famous saying “research is like a step function”, I am always positive and full of hope in front of my research. Meanwhile, I would like to show my deepest gratitude to my co-advisor, Associate Prof. Ling Chen. She helped me revise my papers, invited me good meals, and often encouraged me, even when I felt very frustrated in my early research. Without their extraordinary support and supervision, I could not achieve what I have right now. I feel extremely grateful for their efforts and friendships.

Moreover, I am greatly indebted to Prof. Chengqi Zhang. Without his endless trust and help, I would never have this valuable opportunity to pursue my PhD research in UTS. Meanwhile, I greatly indebted to the Centre for Artificial Intelligence (CAI) directed by Prof. Jie Lu. In CAI, I got tremendous opportunities to learn from many world-famous experts; I met many colleagues with great enthusiasm for scientific research; and I was also permitted to attend many prestigious international conferences related to my research. Studying in UTS and CAI will be a fantastic memory that I will never forget.

During my PhD period, I am very fortunate to join RIKEN-AIP as a research intern in 1-year “AI Residency” Program, working with Prof. Masashi Sugiyama and Dr. Gang Niu. Prof. Masashi Sugiyama always supported my research ideas, commented all my papers carefully, and approved most of my academic travels financially. Dr. Gang Niu often

brought me excellent ideas, denoted the bar of top-tier research, and handed on many technical details. Our research styles matched very well. Meanwhile, Prof. Mingyuan Zhou from University of Texas Austin taught me Bayesian machine learning; Prof. Xiaokui Xiao from National University of Singapore taught me differential privacy; Dr. Quanming Yao from 4Paradigm Inc. taught me how to write and revise a good paper, and his work ethic won all my respects; Weihua Hu from Stanford University taught me how to evaluate a good idea, and how to follow excellence.

During my PhD period, I have been extremely fortunate to collaborate with three great minds, Jiangchao Yao, Yuangang Pan, and Xingrui Yu, who have prominent impact on my research with their outstanding expertise and profound vision. I will never forget the collaboration experiences with them for catching each deadline. Meanwhile, I must thank to all my wonderful friends, classmates and colleagues, Prof. Yi Yang, Dr. Guodong Long, Dr. Miao Xu, Dr. Liang Zheng, Dr. Zhongwen Xu, Dr. Yan Yan, Dr. Mingming Gong, Dr. Tongliang Liu, Dr. Chen Gong, Dr. Xiyu Yu, Dr. Guoliang Kang, Dr. Shaoli Huang, Dr. Liu Liu, Dr. Jiang Bian, Dr. Jing Chai, Dr. Xiaobo Shen, Dr. Haishuai Wang, Dr. Xun Yang, Dr. Guansong Pan, Dr. Peng Hao, Dr. Qinzhe Zhang, Dr. Fan Zhang, Yiliao Song, Feng Liu, Hehe Fan, Hu Zhang, Yanbing Liu, Linchao Zhu, Pingbo Pan, Zhedong Zheng, Ruiqi Hu, Donna Xu, Yaxin Shi, Bo-Jian Hou, Yao-Xiang Ding, and many others, for every enjoyable moment.

Last but not the least, my gratitude also extends to all my family members who have been consistently supporting, encouraging and caring for me all of my life! Most importantly, I feel strongly indebted to my parents Junxiao Han and Jianping Wang. I would never have gone so far without their unconditional love, accompany and support. I would also express my special thanks to my aunt Jianhua Wang, uncle Zhisheng Si and cousin Wei Si for their great help in many years.

Abstract

Modern machine learning is migrating to the era of complex models (i.e., deep neural networks), which requires a plethora of well-annotated data. Crowdsourcing is a promising tool to achieve this goal, since a plethora of labels that can be efficiently collected from crowdsourcing services at very low cost. However, existing crowdsourcing approaches barely acquire a sufficient amount of high-quality labels. This brings the *first* question: How to design the robust mechanism to improve the label quality?

Without such robust mechanism, labels annotated by crowdsourced workers are often noisy, which inevitably degrades the performance of large-scale optimizations, including the prevalent stochastic gradient descent (SGD). Specifically, these noisy labels adversely affect updates of the primal variable in conventional SGD. This brings the *second* question: How to optimize the training model robustly under noisy labels?

Without such robust optimization, it is challenging to train deep neural networks robustly with noisy labels, as the learning capacity of deep neural networks is so high that they can totally memorize and over-fit on these noisy labels. This brings the *third* question: How to acquire the robust model with good generalization under noisy labels? Therefore, in this thesis, we aim to develop a series of robust machine learning approaches, so that they can perfectly handle the difficult from noisy supervision. Our works are summarized as follows:

Chapter 2 answers the first question. Motivated by the “Guess-with-Hints” answer strategy from the Millionaire game show, we introduce the hint-guided approach into crowdsourcing to deal with this challenge. Our approach encourages workers to get help from hints when they are unsure of questions. Specifically, we propose a hybrid-stage setting, consisting of the main stage and the hint stage. When workers face any uncertain question on the main stage, they are allowed to enter the hint stage and look up hints before making any answer. A unique payment mechanism that meets two important design principles is developed. Besides, the proposed mechanism further encourages high-quality workers less using hints, which helps identify and assigns larger possible payment to them.

Experiments are performed on Amazon Mechanical Turk, which show that our approach ensures a sufficient number of high-quality labels with low expenditure and detects high-quality workers.

Chapter 3 answers the second question. We propose a robust SGD mechanism called PrOgressive STochAstic Learning (POSTAL), which naturally integrates the learning regime of curriculum learning (CL) with the update process of vanilla SGD. Our inspiration comes from the progressive learning process of CL, namely learning from “easy” tasks to “complex” tasks. Through the robust learning process of CL, POSTAL aims to yield robust updates of the primal variable on an ordered label sequence, namely from “reliable” labels to “noisy” labels. To realize POSTAL mechanism, we design a cluster of “screening losses”, which sorts all labels from the reliable region to the noisy region. We derive the convergence rate of POSTAL realized by screening losses. Meanwhile, we provide the robustness analysis of representative screening losses. Experiments on benchmark datasets show that POSTAL using screening losses is more effective and robust than several existing baselines.

Chapter 4 answers the third question. Motivated by the memorization effects of deep networks, which shows networks fit clean instances first and then noisy ones, we present a new paradigm called “*Co-teaching*” combating with noisy labels. We train two networks simultaneously. First, in each mini-batch data, each network filters noisy instances based on memorization effects. Then, it teaches the remained instances as the useful knowledge to its peer network for updating the parameters. Empirical results on three benchmark datasets demonstrate that, the robustness of deep learning models trained by Co-teaching approach is much superior than that of state-of-the-art methods.

Contents

Contents	vii
List of Figures	x
List of Tables	xiii
1 Introduction	1
1.1 Motivations	1
1.2 Background	3
1.2.1 Mechanism	3
1.2.2 Optimization	4
1.2.3 Generalization	5
1.2.4 Real-world Applications	5
1.3 Related Work	6
1.3.1 Robust Mechanism	6
1.3.2 Robust Optimization	8
1.3.3 Robust Generalization	10
1.4 Contributions	11
1.5 Thesis Structure	15
1.6 Publications during PhD Study	15
2 Millionaire: Towards Robust Mechanism	18
2.1 Problem Setup	19
2.1.1 Hint-guided Approach	19
2.1.1.1 Hybrid-stage Setting	19
2.1.1.2 Model Assumption	21
2.1.1.3 Payment Mechanism	22
2.1.1.4 Difference from Previous Approaches	23
2.1.2 General Rules of Hints	23
2.1.3 Needs of Hybrid-stage Setting	24
2.2 Hint-guided Payment Mechanism	24

2.2.1	Design Principles	25
2.2.2	Proposed Payment Mechanism	25
2.2.3	No Other Compatible Mechanism	33
2.3	Numerical Experiments	34
2.3.1	Experimental Setup	34
2.3.2	Payment Parameters	36
2.3.3	Experimental Results	37
2.3.3.1	Label Quantity	37
2.3.3.2	Label Quality	38
2.3.3.3	Worker Quality Dectection	38
2.3.3.4	Spammer Prevention	39
2.3.3.5	Money Cost	39
2.4	Summary of This Chapter	40
3	POSTAL: Towards Robust Optimization	43
3.1	Mechanism of Progressive Stochastic Learning	44
3.2	Realization of Progressive Stochastic Learning	47
3.3	Analysis of Progressive Stochastic Learning	50
3.3.1	Convergence Analysis	50
3.3.2	Robustness Analysis	52
3.4	Experiments	55
3.4.1	Experimental Setup	56
3.4.1.1	Baselines	56
3.4.1.2	Parameters	57
3.4.2	Empirical Study on UCI Simulated Datasets	57
3.4.2.1	Effectiveness of POSTAL Mechanism	57
3.4.2.2	Effectiveness of POSTAL using Screening Losses	62
3.4.2.3	Robustness Improvement	62
3.4.3	Real-World Application on AMT Crowdsourced Data	63
3.4.3.1	Robustness Improvement	63
3.5	Summary of This Chapter	64
4	Co-teaching: Towards Robust Generalization	65
4.1	Co-teaching Meets Noisy Supervision	66
4.1.1	Two important questions	66
4.1.2	Relations to Co-training	68
4.2	Experimental Setup	68
4.2.1	Datasets	68
4.2.2	Baselines	70
4.2.3	Network Structure	71
4.2.4	Noise Rates	72

CONTENTS

4.2.5	Measurements	72
4.3	Empirical Results	72
4.3.1	Results on <i>MNIST</i>	72
4.3.2	Results on <i>CIFAR10</i>	74
4.3.3	Results on <i>CIFAR100</i>	75
4.3.4	Choices of $R(T)$	75
4.3.5	Choices of τ	76
4.4	Full Figures	77
4.4.1	<i>MNIST</i>	78
4.4.2	<i>CIFAR10</i>	79
4.4.3	<i>CIFAR100</i>	80
4.5	Summary of This Chapter	80
5	Conclusion and Future Work	81
5.1	Thesis Summarization	81
5.2	Future Work	82
	References	84

List of Figures

1.1	A task that requires workers to answer the question “Which one is the Sydney Harbour Bridge?”. Top panel: the proposed interface under the hybrid-stage setting, consists of two options (“A” and “B”) and a “? & Hints” button. Bottom panel: when workers feel unsure of this question and click the button, the content of hints (gray) is visible, which guides workers to make a choice.	13
1.2	Comparison of error flow among MentorNet (M-Net) [Jiang et al., 2018], Decoupling [Malach and Shalev-Shwartz, 2017] and Co-teaching. Assume that the error flow comes from the biased selection of training instances, and error flow from network A or network B network is denoted by red arrows or blue arrows, respectively. Left panel: M-Net maintains only one network (A). Middle panel: Decoupling maintains two networks (A & B). The parameters of two networks are updated, when the predictions of them disagree (!=). Right panel: Co-teaching maintains two networks (A & B) simultaneously. In each mini-batch data, each network samples its small-loss instances as the useful knowledge, and teaches such useful instances to its peer network for the further training. Thus, the error flow in Co-teaching displays the zigzag shape.	14
1.3	The structure of this thesis.	16
2.1	Mathematical model of the decision process under our hybrid-stage setting.	21
2.2	Evaluation of label quality. Percentage (in %) of correct answers and incorrect answers on three tasks are provided. Note that, we do not plot the percentage of unlabeled questions.	41
2.3	Evaluation of the label quality. Results on <i>Speech Clips</i> are not reported, as text answers cannot be majority voted.	42
2.4	Evaluation of the spammer prevention. Average payment to each worker on all three tasks are provided. Evaluation of the money cost. .	42

3.1	Left Panel: Circles denote <i>real positive</i> instances such as “A”. Squares represent <i>real negative</i> instances such as “B” and “C”; however, both “B” and “C” are <i>erroneously</i> annotated as positive class, which creates two noisy labels. According to their distance to the positive hyperplane, the label of “A” is reliable; while the labels of “B” and “C” are noisy. Right Panel: “A” (z_A) is located in the reliable region defined by $z \geq 0$; while “B” (z_B) and “C” (z_C) are located in the noisy region defined by $z \leq 0$. In Definition 3.1, $z = yf_w(\mathbf{x})$ is formally defined as the curriculum in POSTAL. There are two points to be noted that: 1) Ramp Loss is parameterized by s^* directly; and 2) Logistic, Hinge and Ramp Losses are drawn as the baselines of two screening losses.	45
3.2	To verify the <i>effectiveness</i> of POSTAL mechanism, we compare POSTAL with vanilla SGD under the same screening losses. We provide the testing error rate (in %, TER) with the number of epochs on representative UCI noisy datasets: small-scale <i>A7A</i> , large-scale <i>SUSY</i> and high-dimensional <i>REAL-SIM</i> with 20% and 40% of noisy labels. .	58
3.3	To verify the <i>effectiveness</i> of POSTAL using screening losses, we compare POSTAL(SRamp) and POSTAL(SGomp) with stochastic methods in the second category of baselines (paragraph 1 in Section 3.4.1). We provide the testing error rate (in %, TER) with the number of epochs on representative UCI noisy datasets: small-scale <i>A7A</i> , large-scale <i>SUSY</i> and high-dimensional <i>REAL-SIM</i> with 20% and 40% of noisy labels. For PEGASOS, the number of epochs for convergence is set to $\frac{10}{\lambda} (\gg 15)$ by default. Thus, its result is not reported.	59
3.4	To verify the <i>robustness</i> of POSTAL using screening losses, we compare POSTAL(SRamp) and POSTAL(SGomp) with all baselines on all UCI noisy datasets. We provide the testing error rate (in %, TER) with varying percentages (0% to 40%) of noisy labels. Note that the color of each method is uniform.	60
3.5	To verify the <i>robustness</i> of POSTAL using screening losses, we compare POSTAL(SRamp) and POSTAL(SGomp) with all baselines on all UCI noisy datasets. We provide the variance with varying percentages (0% to 40%) of noisy labels. Note that the color of each method is uniform.	61
3.6	Sample images of four categories of dogs from the <i>Stanford Dogs</i> dataset.	63
4.1	Transition matrices of different noise types (using 5 classes as an example).	69
4.2	Test accuracy vs number of epochs on <i>MNIST</i> dataset.	73
4.3	Label precision vs number of epochs on <i>MNIST</i> dataset.	74

LIST OF FIGURES

4.4	Results on <i>CIFAR10</i> dataset. Top: test accuracy vs number of epochs; bottom: label precision vs number of epochs.	75
4.5	Results on <i>CIFAR100</i> dataset. Top: test accuracy vs number of epochs; bottom: label precision vs number of epochs.	77
4.6	Results on <i>MNIST</i> dataset. Top: test accuracy vs number of epochs; bottom: label precision vs number of epochs.	78
4.7	Results on <i>CIFAR10</i> dataset. Top: test accuracy vs number of epochs; bottom: label precision vs number of epochs.	79
4.8	Results on <i>CIFAR100</i> dataset. Top: test accuracy vs number of epochs; bottom: label precision vs number of epochs.	80

List of Tables

2.1	Comparison of related approaches and our hint-guided approach (Chapter 2). Baseline is the approach explained above. The skip-based approach comes from [Ding and Zhou, 2017; Shah and Zhou, 2015]. Note that, the self-corrected approach [Shah and Zhou, 2016] is designed theoretically, but barely realized for real tasks (denoted as “×”), so its metrics of “label quality” and “money cost” cannot be evaluated (denoted as “-”). However, since its payment mechanism has a multiplicative form, it prevents spammers theoretically.	18
2.2	The payment parameters. The total payment is composed of FP and RP . RP is based on worker’s responses to G questions. $P_h = \frac{\frac{1}{2}-\epsilon}{2T-1}P_d$ according to Algorithm 1, where we set $T = 0.75$ and $\epsilon = 0.191$ due to the Proposition 2.	36
2.3	Evaluation of the label quantity. We provide the percentage of the completion on three tasks.	38
2.4	Evaluation of the worker quality detection of the hint-guided approach. Error rate (in %) is provided for aggregating original and rescaled crowdsourced labels.	39
3.1	Comparison of different learning approaches. POSTAL integrates benefits (emphasized by color) of CL and SGD.	44
3.2	Datasets used in the UCI simulated study. Representative datasets are emphasized by color.	56
3.3	To verify the robustness of POSTAL using screening losses, we also provide the comparison of “testing error rate (in %) with standard deviation” among all methods on UCI clean datasets. Methods indicated by “-”run out of memory.	58

3.4	To verify the robustness of POSTAL using screening losses in real-world situations, we provide the comparison of “testing error rate (in %, TER) with standard deviation” on four crowdsourcing subsets. In each subset, the ratio between positive samples and negative samples is around 1 : 1. For LIBLINEAR, the left TER is from LIBPrimal while the right TER is from LIBDual.	64
4.1	Summary of data sets used in the experiments.	68
4.2	Comparison of state-of-the-art techniques with our Co-teaching approach. In the first column, “large noise”: can deal with a large number of class; “heavy noise”: can combat the heavy noise, i.e., high noise ratio; “flexibility”: need not combine with specific network architecture; “no pre-train”: can be train from scratch.	70
4.3	CNN and MLP models used in our experiments on <i>MNIST</i> , <i>CIFAR10</i> , and <i>CIFAR100</i> . The slopes of all LReLU functions in the networks are set to 0.01.	71
4.4	Average test accuracy on <i>MNIST</i> over the last ten epoch.	73
4.5	Average test accuracy on <i>CIFAR10</i> over the last ten epoch.	74
4.6	Average test accuracy on <i>CIFAR100</i> over the last ten epoch.	76
4.7	Average test accuracy on <i>MNIST</i> over the last ten epoch.	76
4.8	Average test accuracy of Co-teaching with different τ on <i>MNIST</i> over the last ten epoch.	77