

# **Distribution-based Active Learning**

by

**Xiaofeng Cao**

Australian Artificial Intelligence Institute  
University of Technology Sydney

Supervisor: Prof. Ivor W. Tsang  
A thesis submitted for the degree of

*Doctor of Philosophy*

Sydney, Australia

2021

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

This research is supported by the Australian Government Research Training Program under Grant DP180100106 and DP200101328.

Production Note:

Signature: Signature removed prior to publication.

Date: *30-Apr-2021*

## Acknowledgements

In a sense, Prof. Ivor W. Tsang brought hope for my research career. The early journey of pursuing the Ph.D. degree in Australia was tortuous and arduous. As the Hong Kong film propagated in my childhood, “*professional*” is the most important impression Ivor gave me. He gradually reversed my dilemma on the academic path using his values. He is a humble scholar who maintains earnest and prudent attitudes, leading me to be a righteous person. I can never forget what he often told me: “*what kind of person you want to be, what you will do*”.

My research went smoothly when I was a master’s degree candidate in China. This leads me to ambitiously select some challenging research topics at the beginning of my Ph.D. study. Active learning theory, a fundamental topic in PAC learning, started my Ph.D. journal. Also because of this, it is difficult to understand those theorems and proofs at the beginning, since my master research background was totally about applications. However, Ivor never gave up on me. Continuous encouragement and enthusiasm inspired me to explore the deeper theoretical mysteries behind those complicated formulas and theorems. My theoretical ability thus has been greatly improved. This is something I never thought of. No matter how difficult the time, I always insisted on one truth: “*if you cannot please others, you can please yourself*”. This made me gradually understand why Ph.D. was termed philosophy in early western culture. The harvested values will accompany me all my life.

I met my wife in Australia. Lin is my biggest support in this learning journey. Love and marriage can be selfless. We married at 160 Marsden St, BDM Parramatta, Sydney. Thanks to the wedding witness officer Ms. Aila and my friend Mr. Shuai Cheng. Thanks my co-supervisor Prof. Ling Chen for her help in my study. I want to thank my father-in-law and mother-in-law. They provided strong support in my study whether in funding or life. This allows us to establish a small family at 174/11 Potter Street, Waterloo, Sydney, where provides us with very high-quality study and life. I know they do not care about these expenses. However, it greatly relieved my stress and anxiety during my second half of pursuing the PhD degree. My father-in-law is a farsighted person in Hong Kong. I saw the elite spirits of the Cantonese and Chaoshan. His extraordinary achievements inspired me to work harder and go further. My mother-in-law is an elegant lady who requires Lin to unconditionally support my study in daily life. I want to thank my parents, brothers, and sister in law for their consistent support in my study. Specially thanks to my brother, he supported our family and relieved my stress during this period.

*“Be humble, be reasonable, and be respected.”*

—It is the advice from Ivor before I am leaving UTS and going to start my personal academic career. I will spread this motto to the younger students.

Xiaofeng Cao  
Sydney, Australia, 2021.

## Abstract

Active learning aims to maximize the learning performance of the current hypothesis by drawing as few labels as possible from an input distribution. To build a near-optimal hypothesis, halfspace learning improved the generalization of a perceptron vector over a unit sphere, presenting model guarantees for the reliable (practical) active learning, in which the error disagreement coefficient controls the hypothesis update via pruning the hypothesis class. However, this update process critically depends on the initial hypothesis and the coefficient. Their improper settings may improve the bounds on the label complexity, which estimates the label demands before achieving a desired error for the hypothesis. One question thus arises: how to reduce the label complexity bounds? In a worse situation, estimating updates of hypothesis using error lacks feasible guarantees, if the initial hypothesis is a null (insignificant) hypothesis. Another question also arises: how to control the hypothesis update without errors, when estimating the error disagreement is infeasible? For error disagreement, most of its generalizations regarding to hypothesis update, either make strong distribution assumptions such as halfspace learning, or else they are computationally prohibitive. How to improve the performance of deep active learning based on the theoretical results of active learning of halfspace?

This thesis tries to answer the three questions from shattering, disagreeing, and matching over distributions. With halfspace learning, the first work presents a novel perspective of shattering the input distribution that, guaranteeing from a lower bound on Vapnik-Chervonenkis (VC) dimension, further reduces the label complexity of active learning. When estimating errors is infeasible, the second work proposes a distribution disagreement graph coefficient, which estimates hypothesis from distribution, yielding a tighter bound on typical label complexity. The constructed hyperbolic model, generalizing distribution disagreement by focal representation, shows effective improvements compared to generalization algorithms of error disagreement. On deep learning settings for active learning, the Bayesian neural network shows expressive distribution matching on the massive training parameters, which allows estimating error disagreement can work effectively. We thus integrate the error and distribution disagreements to establish a uniform framework, which matches the geometric core-set expression of the distribution, interacting with a deep learning model.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Motivation . . . . .	1
1.2	Background . . . . .	3
1.2.1	Active Learning of Halfspace . . . . .	3
1.2.2	Error Disagreement . . . . .	5
1.3	Related Work . . . . .	7
1.3.1	Halfspace Learning . . . . .	7
1.3.2	Hyperbolic Geometry . . . . .	8
1.3.3	Deep Active Learning . . . . .	8
1.4	Contributions . . . . .	9
1.4.1	Tighter Label Complexity Bounds . . . . .	9
1.4.2	Distribution Disagreement . . . . .	10
1.4.3	A Unified Framework . . . . .	11
1.5	Thesis Organizations . . . . .	12
1.6	Publications . . . . .	13
<b>2</b>	<b>Distribution Shattering</b>	<b>15</b>
2.1	Main Theoretical Insights . . . . .	15
2.1.1	Error Disagreement of Hypothesis-pruning . . . . .	16
2.1.2	Sparse Hypothesis Class . . . . .	17
2.1.3	Performance Disagreement . . . . .	18
2.2	Hypothesis and Distribution . . . . .	20
2.2.1	Monotonic Property of the Active Query Set . . . . .	20
2.2.2	Error Rate Change by Shattering Number Density . . . . .	20
2.2.3	Bottleneck of Hypothesis-pruning . . . . .	21
2.3	Distribution-shattering for Active Learning . . . . .	22
2.3.1	Shattering: From Halving to Splitting . . . . .	22
2.3.2	Halving Number Density for Shattered Distribution . . . . .	22
2.3.3	Advantages of Shattered Distribution . . . . .	23
2.3.3.1	Bounds of Error Disagreement in Shattered Distribution . . . . .	23
2.3.3.2	Fall-back Analysis in Shattered Distribution . . . . .	24
2.3.3.3	Bounded Noise Analysis of Shattered Distribution . . . . .	24
2.3.4	Distribution-shattering for Active Learning Tasks . . . . .	25
2.3.4.1	Active Scoring for Halving . . . . .	25
2.3.4.2	Splitting by Distribution Density . . . . .	26
2.3.4.3	Querying by SDAL . . . . .	26
2.4	Experiments . . . . .	27

2.4.1	Experimental Setup . . . . .	27
2.4.2	Effectiveness of Halving . . . . .	29
2.4.3	Optimal Error of Querying . . . . .	29
2.4.4	Average Error of Querying . . . . .	31
2.4.5	Querying with Adversarial Examples . . . . .	32
2.4.6	Querying with Noisy Labels . . . . .	35
2.4.7	Calculation Complexity . . . . .	36
2.5	Discussions . . . . .	36
2.5.1	Distribution-shattering and Experimental Design . . . . .	36
2.5.2	Learning Curves of Sequential Sampling in Shattered Space . . . . .	39
2.6	Summary of This Chapter . . . . .	39
<b>3</b>	<b>Distribution Disagreeing</b>	<b>41</b>
3.1	Distribution Disagreement Graph Coefficient . . . . .	41
3.1.1	Graph Coefficient . . . . .	41
3.1.2	Tightness of Approximation . . . . .	45
3.1.3	Why Hyperbolic Geometry? . . . . .	47
3.2	Hyperbolic Focal Representation . . . . .	50
3.2.1	Squared Lorentzian Distance . . . . .	50
3.2.2	Geometric Centroids . . . . .	51
3.2.3	Lorentzian Focal Point Approximation . . . . .	52
3.2.4	Lorentzian Focal Representation . . . . .	53
3.2.5	Tree-likeness Splitting . . . . .	54
3.3	Experiments . . . . .	56
3.3.1	Experimental Setup . . . . .	56
3.3.2	Centroid vs. Focal Representations . . . . .	57
3.3.3	Deep Active Learning with ResNet20 . . . . .	58
3.3.4	Learning Curves of ResNet20 . . . . .	59
3.3.5	Visualization of Lorentzian Focal Representations . . . . .	60
3.3.6	Batch Performance of Deep Active Learning . . . . .	62
3.4	Discussions . . . . .	62
3.4.1	Deep Active Learning with Weak-supervision . . . . .	62
3.4.2	Limitations of Our Approach . . . . .	63
3.4.3	Computational Complexities . . . . .	63
3.5	Summary of This Chapter . . . . .	64
<b>4</b>	<b>Distribution Matching</b>	<b>65</b>
4.1	Bayesian Active Learning by Disagreements . . . . .	65
4.1.1	Interpretations . . . . .	66
4.1.1.1	Model Uncertainty Estimation . . . . .	67
4.1.1.2	Core-set Construction . . . . .	67
4.2	Framework . . . . .	68
4.2.1	Geometric Interpretation of Core-set . . . . .	68
4.2.2	Initial Acquisitions based on Core-set Construction . . . . .	69
4.2.3	Model Uncertainty Estimation with Core-set . . . . .	71
4.2.4	Two-stage GBALD Algorithm . . . . .	72

4.3	Experiments . . . . .	72
4.3.1	Experimental Setup . . . . .	72
4.3.2	Uninformative Priors . . . . .	75
4.3.3	Improved Informative Acquisitions . . . . .	75
4.3.4	Active Acquisitions . . . . .	76
4.3.5	Active Acquisitions with Repeated Samples . . . . .	79
4.3.6	Active Acquisitions with Noisy Samples . . . . .	79
4.3.7	BatchBALD vs. GBALD . . . . .	80
4.4	Discussions . . . . .	81
4.4.1	Acceleration of Accuracy . . . . .	81
4.4.2	Hyperparameter Settings . . . . .	82
4.4.3	Two-sided $t$ -test . . . . .	84
4.5	Summary of This Chapter . . . . .	85
<b>5</b>	<b>Conclusion</b>	<b>86</b>
5.1	Thesis Summarization . . . . .	86
5.2	Future Work . . . . .	86
	<b>References</b>	<b>88</b>
	<b>Appendix</b>	<b>97</b>
	Proofs of Chapter 2 . . . . .	97
	Proofs of Chapter 3 . . . . .	98

# List of Figures

1.1	Active learning queries unlabeled data selected by the machine learning model. The querying loops stop until the model achieved a desired generalization performance. . . . .	2
1.2	Halfspace learning over a unit sphere with a radius of $R$ , where $+$ , $-$ denote different class labels. The error disagreement-based active learning strategy prunes the hypothesis set (reduce the number of candidate hypotheses, i.e., the diameters across the colored regions) via querying data distributed in the colored pool. . . . .	4
1.3	Halfspace learning presents guarantees for the theoretical analysis and generalization performance of distribution-based active learning. Theoretical analysis includes contributions of tighter label complexity bounds of Chapter 1.4.1 and distribution disagreement graph coefficient of Chapter 1.4.2. Generalization performance includes hyperbolic focal representation which generalizes distribution disagreement in hyperbolic geometry (related work in Chapter 1.3.2), and deep active learning framework of Chapter 1.4.3 (related work in Chapter 1.3.3). . . . .	7
1.4	A uniform framework consists generalizations of distribution and error disagreements. . . . .	11
1.5	The skeleton of the thesis. . . . .	12
2.1	Error disagreement of halfspace learning over a unit sphere with a radius of $R$ , where $+$ and $-$ symbols denote the positive and negative class labels, respectively. The error disagreement-based active learning strategy prunes the hypothesis set via updating a subsequent hypothesis with large $\vartheta$ ; the initial hypothesis has an error of $1/2$ , and the subsequent hypothesis has an error of $\vartheta/\pi$ , thereby the error disagreement is $1/2 - \vartheta/\pi$ . . . . .	16
2.2	Shattering distribution of a unit sphere with a radius of $R$ , where $+$ , $-$ denote the class labels. Distribution-shattering halves the number density of an input distribution w.r.t. $\frac{12}{\pi R^2}$ into a shattered distribution w.r.t. $\frac{6}{\pi R^2}$ . Any hypothesis generated from the original distribution is characterized with a lower bound on VC dimension. Thereby, we can find a representation structure that induces a tighter label complexity without estimating the hypothesis. . . . .	18



2.3	Number density models the relationship of hypothesis and distribution, i.e. VC dimension increases proportionally as the number density. If number density= $\frac{4}{a^3}$ and the hypothesis is generated from $i$ samples, the VC bound $\leq \sum_{i=1}^4 \binom{4}{i} = 16$ ; given the number density= $\frac{6}{a^3}$ , the VC bound $\leq \sum_{i=1}^6 \binom{6}{i} = 64$ . Note the version space of number density= $\frac{6}{a^3}$ only presents 32 hypotheses, where the red nodes are the newly added ones from the version space of number density= $\frac{4}{a^3}$ , and each blue node denotes one feasible hypothesis, respectively. . . . .	19
2.4	The assumption of distribution-shattering with a sparse hypothesis class. Each node denotes one realizable hypothesis, and the lengths of the red lines denote the hypothesis disagreement. Hypothesis-pruning updates the initial hypothesis w.r.t. $h_0$ into $h_\epsilon$ with a desired error $\epsilon$ in the original hypothesis class (Fig. 2.4(a)). Distribution-shattering optimizes a group of local hypothesis spheres (Fig. 2.4(b)). Shattering by those sphere centers, the original hypotheses are transformed into a sparse hypothesis class (Fig. 2.4(c), thereby finding $h_\epsilon$ can achieve a lower label complexity than in its original hypotheses. . . . .	21
2.5	Error rate changes of undertaking passive sampling in input and shattered spaces on different datasets. . . . .	29
2.6	The error rate performances of the seven active learning approaches on the active learning test. (a)-(d) are four UCI datasets; (e)-(k) are that on the selected sub datasets of <i>letter</i> , where the class number of them are 12, 16, 20, and 26. . . . .	30
2.7	Illustration of the produced adversarial examples by FGSM with different perturbation parameters, where the marked examples are clean data without feature perturbations. . . . .	33
2.8	The performance of error rate on active learning querying with adversarial examples, where the adversarial examples are produced by FSGM with different perturbation parameter settings. . . . .	34
2.9	The performance of error rate on active learning querying with noisy labels. . . . .	36
2.10	Representative samples of MAED using different $\lambda$ . . . . .	37
2.11	Representative samples of SDAL using different initializations ( $r = 0.25$ , $\epsilon = 0.1$ ). . . . .	38
2.12	Error rate changes of undertaking sequential sampling in the shattered space and original input space. . . . .	39
3.1	Motivation of distribution disagreement graph coefficient over halfspace learning. As the decrease of number density, the graphs covering all feasible hypotheses keep consistent structures, where each vertex denotes one hypothesis and each edge denotes the hypothesis disagreement (distance) of the two connected vertices. Learning a halfspace with graph coefficient can completely replace the error disagreement. . . . .	42

	10
3.2	Generalization test over MNIST dataset using $\theta$ and $\theta_G$ . . . . . 46
3.3	Euclidean, Gaussian kernelized, Poincaré and Lorentzian centroids on a noisy spherical Gaussian dataset. The minimization or maximization on the centroids are performed with 30 iterations against $k$ -medoids algorithm. . . . . 48
3.4	Test accuracies of Euclidean, squared Euclidean, Gaussian kernelized, Poincaré and Lorentzian centroids with varying parameters. . . . . 49
3.5	The Lorentzian version space ( $\nabla A'BC, \nabla D'EF$ ) vs. Euclidean version space ( $\nabla ABC, \nabla DEF$ ). Lorentzian norm shrinks the volume of Euclidean version space by shifting Euclidean centroids (A,D) into Lorentzian focal points ( $A', D'$ ) that are close to the boundary region. As such, sampling from regions $ABA'C$ and $DED'F$ is ineffective, which means any active learning model in the Lorentzian version space has a tighter bound on label complexity. . . . . 50
3.6	Geometric centroids of Euclidean, Poincaré, and focal points of squared Lorentzian distances in a given set of points. Lorentzian norm updates the focal points of the embedded half-sphere toward the surface as the parameter $\mathcal{B}$ decreases w.r.t. Eq. (3.14). . . . . 53
3.7	Illustration of tree-likeness splitting. The first layer nodes are $k$ global focal points. A binary tree splitting strategy begins from the second layer of the tree. . . . . 55
3.8	Test accuracies of Euclidean, squared Euclidean, Gaussian kernelized, Poincaré centroids and our Lorentzian focal points with varying parameters on MNIST. . . . . 58
3.9	Test accuracies of training ResNet20 with different active learning outputs on CIFAR-10, CIFAR-100, and SVHN. . . . . 58
3.10	Test accuracies of each epoch of training ResNet20 with different active learning outputs on CIFAR-10, CIFAR-100, and SVHN. (Before 80 epochs, the learning rate is 0.001, and afterwards it is 0.1.) . . . . . 60
3.11	Training losses of each epoch of training ResNet20 with different active learning outputs on CIFAR-10, CIFAR-100, and SVHN. (Before 80 epochs, the learning rate is 0.001, and afterwards it is 0.1.) . . . . . 60
3.12	Projection of Lorentzian focal points in the first two dimensional feature space of CIFAR-10. . . . . 61
3.13	Lorentzian focal vs. batch performance of error entropy and error variation on CIFAR-10. . . . . 62
4.1	Illustration of two-stage GBALD framework. BALD has two types of expression: model uncertainty estimation and core-set construction where the deeper the color of the core-set elements, the higher the representation; GBALD integrates them to an uniform framework. Stage ①: core-set construction is with an ellipsoid, not typical sphere, representing the original distribution to indicate the input features of DNN. Stage ②: model uncertainty estimation of DNN then derives the subsequent highly informative and representative samples. . . . . 66

4.2	Optimizing BALD with sphere and ellipsoid geodesics. Ellipsoid geodesic rescales the sphere geodesic to prevent the updates of the core-set towards the boundary regions of the sphere where the characteristics of the distribution cannot be captured. Black points denote the feasible updates of the red points. Dash lines denote the geodesics. . .	68
4.3	Acquisitions with uninformative priors from digit ‘0’ and ‘1’. . . . .	74
4.4	GBALD outperforms BALD using ranked informative acquisitions which cooperate with representation constraints. . . . .	76
4.5	Active acquisitions on MNIST, SVHN, and CIFAR10 datasets. . . . .	77
4.6	Active acquisitions on SVHN with 5,000 and 10,000 repeated samples. .	78
4.7	Active noisy acquisitions on SVHN with 5,000 and 10,000 noisy labels. .	80
4.8	Comparisons of BALD, BatchBALD, and GBALD of active acquisitions on MNIST with bath settings. . . . .	81
4.9	Accelerations of accuracy of different baselines on MNIST, SVHN, and CIFAR10 datasets. . . . .	82
4.10	Accelerations of accuracy of active acquisitions on SVHN with 5,000 and 10,000 repeated samples. . . . .	82
4.11	Accelerations of accuracy of active noisy acquisitions on SVHN with 5,000 and 10,000 noisy labels. . . . .	83
4.12	A pairwise penalty matrix over active acquisitions on MNIST, SVHN, and CIFAR10. Column-wise values at the bottom of each matrix show the overall performance of the compared baselines (larger value has more significant superior performance). . . . .	85

# List of Tables

2.1	The statistical results (mean±std in %) of error of different active learning baselines on six real-world datasets . . . . .	32
2.2	The statistical results (mean±std in %) of active learning with adversarial examples of different algorithms . . . . .	35
2.3	The statistical results (mean±std in %) of error of experimental design active learning baselines and SDAL on three real-world datasets . . . . .	37
3.1	Mean±standard deviation of the breakpoints of the generalization test over MNIST using $\theta$ and $\theta_G$ . . . . .	46
3.2	Label complexities of the generalization test over MNIST using $\theta$ and $\theta_G$ (Radius denotes “radius of hypothesis class”) . . . . .	47
3.3	Summary of distance metrics and centroid expressions in the Euclidean space, RKHS, and hyperbolic space. . . . .	48
3.4	Mean±standard deviation of the test accuracies on CIFAR-10, CIFAR-100 and SVHN after 80 epochs over 3 runs. . . . .	61
3.5	Mean±standard deviation of the training losses on CIFAR-10, CIFAR-100 and SVHN after 80 epochs over 3 runs. . . . .	61
3.6	Mean±standard deviation of the test accuracies of different baselines with batch settings on CIFAR-10. . . . .	63
3.7	Time computations of selecting 10,000 examples from CIFAR-10 using different baselines. . . . .	63
4.1	Mean±std of the test accuracies of the breakpoints of the learning curves on MNIST, SVHN, and CIFAR-10. . . . .	77
4.2	Number of acquisitions on MNIST, SVHN and CIFAR10 until 70%, 80%, and 90% accuracies are reached. . . . .	78
4.3	Mean±std of active acquisitions on SVHN with 5,000 and 10,000 repeated samples. . . . .	78
4.4	Mean±std of active noisy acquisitions on SVHN with 5,000 and 10,000 noises. . . . .	80
4.5	Mean±std of BALD, BatchBALD, and GBALD of active acquisitions on MNIST with batch settings. . . . .	81
4.6	Relationship of accuracies and sizes of core-set on SVHN. . . . .	83